nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

How early life shapes the infant gut microbiome and risk of disease

PAGES 583 & 589

SECURITY BLANKET

SPACE

REACHING FOR THE MOON

Plans for a permanent lunar base gather momentum

PAGE 474

NEUTRINO PHYSICS

THE HEART OF THE SUN

Neutrino detector probes solar nuclear reactions

PAGES 496 & 505

CAREERS

SATISFACTION IN SCIENCE?

Nature's salary survey takes the pulse of the workplace

PAGE 611

NATURE.COM

25 October 2018

Vol. 562, No. 7728

THIS WEEK

EDITORIALS

ANNOUNCEMENT Matters Arising replaces Brief Communications Arising **p.460**

WORLD VIEW Set up artificial-intelligence centres in Africa **p.461**

AWARDS High-resolution microscopy among Breakthrough winners **p.465**

Ground truths

Nature *survey shows most scientists are happy at work, but that a significant number still face discrimination — an unacceptable situation.*

There is a tendency when drawing conclusions from survey data to look for specific landmarks that instil confidence in the results. Some are defined by cold mathematics: 71% can be presented as "most people" without much controversy. But many others are subject to interpretation. Do 14 of 16 people constitute "almost everybody", or does that take 15?

Perhaps most important is how to handle the lower reaches. Can a single voice from 100 be written off as an outlier? What about two or three? How big does a minority have to become before it gets a bullet point in a report? The data — the numbers that surveys produce — usually tell the full story to those who are willing to look, but most (that word again) of us rely on a more human narrative to make sense of the results. And here, care is needed.

This week, *Nature* publishes the results of our biennial survey of the income and career satisfaction of scientists across the world. And one narrative that emerges is heartening. Most — 68% — said they were satisfied or very satisfied with their careers. And just over half — 51% — had received a pay rise in the past year.

The majority may rule, but it hardly tells the whole story. Nobody should take any comfort, for example, from the fact that most scientists (72%) told the survey they have not witnessed any instances of harassment or discrimination. The corollary of that figure is clear: 28% have. Nobody can be satisfied with that. The problem of harassment has received some much-needed attention in recent years, but, as these figures show, there remains much work to be done, and attitudes and behaviours still need to change.

The survey has limitations. The results are based on the anonymous responses of 4,334 (self-selected) people who have pursued science beyond an undergraduate degree. Three-quarters of them are based in North America or Europe. Still, many of the figures do mirror those of other surveys — high levels of job satisfaction among scientists working across academia and industry, for instance. Researchers are generally a content and motivated bunch. But look in the margins and there remains much room for improvement.

Poor mental health continues to be a huge concern, with more than one-third (36%) of respondents saying they needed or were receiving help for depression or anxiety. Attitudes from colleagues were not always supportive. "I had a mental health crisis and instead of helping I was suspended from work and threatened with potential dismissal," wrote one. Many universities are aware of this issue and are working to improve care and support. But not all are succeeding.

The survey reveals other institutional failings, too. Sadly, only half of university scientists said their institution was doing enough to promote diversity. And 21% said they had personally experienced harassment or discrimination. This was most commonly based on gender, but the list also included discrimination based on race, religion, sexuality and age. One respondent wrote: "Co-workers have scheduled important meetings on religious holidays and when I object or do not attend, I'm viewed as someone who doesn't take their job seriously." Another said: "A liberal faculty will shun and even harass conservative Christians, mocking them openly."

Some 23% of people who replied to the survey reported discrimination based on age. One respondent complained of "Pressure to retire as I approach age 60. Not explicit or stated, but moral pressure and looks." And about the same number (22%) said they had suffered racial bias.

This is unacceptable. Science must do better on these issues, as individuals and institutions. The survey holds up a mirror to the research community, and if the community does not like what it sees — and it should not — then all of us must do more to change the picture.

"The survey holds up a mirror to the research community."

Science should be a rewarding career. Most scientists say they do enjoy their work and — at least according to this survey — most get through the day without being made to feel that they don't belong, or that they have to do more to prove themselves because of their gender or geographical origin. But "most scientists" here is not enough. Individuals and groups who do experience such abhorrent discrimination must know they are not an overlooked interest. It is everybody's responsibility to condemn such behaviour when they see it. And, where they feel comfortable to do so, everybody should speak out when injustice occurs. ■

Capital thinking

Political attention to human capital must be backed up with solid research.

The surprise 2014 global bestseller *Capital in the Twenty-First Century*, written by French economist Thomas Picketty, highlighted the role of wealth — rather than earnings — in the way that money makes the world go around. But Picketty chose to play down an important part of the system: human capital, the economic value derived from the knowledge, skills and abilities that enable people to perform paid work.

How to include human capital in analyses is as much a political as an economic problem: critics argue that the concept creates a false equivalence between having skills and having money, which plays down financial inequality. Supporters insist that it's a genuine measure of the potential of individuals, populations and nations, and so a way to indicate their intrinsic value.

The World Bank has now reignited the debate. Earlier this month, it released its much-anticipated Human Capital Index (see go.nature.

com/2cwyqqd), which ranks 157 nations according to measures of investment in their people. The bank's measure is relatively simple, constructed from data on child survival and growth, years of primary and secondary schooling, and health. A country can achieve a perfect score if all children born today can expect to survive to 60 without impaired growth and development — resulting from poor nutrition, repeated infection or inadequate psychosocial stimulation, and measured by ratios of height and age — and can expect to have received 14 years of good-quality schooling by age 18.

The index is based on the assumption that a country's economic productivity is tied to the knowledge and abilities of its people. It followed the release, two weeks earlier, of the results of a parallel (but separate) academic exercise by the Institute for Health Metrics and Evaluation at the University of Washington in Seattle (S. S. Lim *et al. Lancet* **392**, 1217–1234; 2018).

The World Bank hopes its new index will mimic the success of its national "ease of doing business" ranking, which has focused government efforts around the world to reduce corruption and encourage outside investment as a way to secure a higher placing than their rivals and competitors. The bank wants to demonstrate how measures of education and health are linked to the productivity and prosperity of a country, assuming that investing in human capital through education and health systems can yield rapid development. In short, it wants to push countries to make things better for their people — and their human capital ranking. It has certainly managed to draw attention: Indian officials immediately protested against their country's low ranking, and government officials there say they will ignore what they argue to be a simplistic and misleading measure.

Top scorers on the World Bank list include Singapore, South Korea and Japan, whereas many African countries, including Mali, Nigeria and Liberia, performed poorly and were near the bottom of the index.

The Institute for Health Metrics and Evaluation based its ranking of 195 countries on similar factors, but incorporated more measures of health and education, and used different data sources and methods. Finland, Iceland and Denmark top its charts, which cover the period from 1990 to 2016. During this time, the United States tumbled from 6th to 27th place, largely owing to minimal progress in educational attainment.

Few would argue against the goal of encouraging better health and education. And perhaps by framing these needs in terms of economic returns and tapping into the political desire to climb the leader boards, these measures might succeed in having a greater impact on decision-makers than do simple appeals to the intrinsic good. For example, one way to improve a country's position would be for it to reduce gender inequality in years of schooling.

"These indices are only as good as the data that underlie them."

But any metric — be it a university ranking or standardized mathematics testing — is selective and must be interpreted appropriately. Too often it becomes a convenient proxy, leading to inferences of quality for which it was never intended, and distorting reality. As in most analyses of this type, these indices are only as good as the data that underlie them. There is a huge range in the quality and quantity of data on both health and education across countries. And although deductions about the exact effects of health outcomes and education on economic productivity are based on research, the true relationships are unclear for the range of countries and contexts to which the Human Capital Index is being applied. Critics are right to point out that a national score does not account for regional differences in a country.

Scientists can play a part here, to ensure that indices such as these become the credible motivators that they are intended to be. More and better data on indicators of health and educational outcomes will improve the accuracy of the indices. More research on rigorous ways to capture other determinants of human capital, and on their relationship to health, prosperity and well-being, will enrich our understanding of how to reach global development goals. *Nature* recognizes the need for such work to help inform policymakers and make their efforts more evidence-based. As such, we encourage submissions of high-quality data and analysis addressing knowledge gaps in assessing and improving human health and well-being. ■

ANNOUNCEMENT

Matters Arising: a venue for commentary

There was a time when scientific progress depended on elaborate and often protracted exchanges of correspondence. Charles Darwin wrote thousands of letters, and his correspondence with influential thinkers had an important impact on his theories. This communication was private. Fortunately, much has survived and found its way into archives, where it forms a key part of the scientific record.

Although research findings today are mainly disseminated and recorded in the form of peer-reviewed research manuscripts, scholarly commentary on published research is still crucial: it can provide nuance, refinement and caveats. And these days, it moves fast.

So, from this week, *Nature* will consider such post-publication contributions as Matters Arising — a format designed to peer-review and publish online exceptionally interesting and timely scientific comments and clarifications related to primary research papers published in the journal. Authors of the original papers will be given the chance to reply. If our editors deem that these responses move the discussion forward in a constructive way, they will be published at the same time as the Matters Arising article.

We also recognize the need for timely release of these exchanges to the relevant communities, and the difficulty of doing so through an often-lengthy peer-review process. So, to accommodate both rigorous peer review and the need for timeliness, authors of Matters Arising and the original *Nature* paper are encouraged to release preprints during the formal journal process, as supported by our policies. Comments can also be made on all original *Nature* research papers online, and these can be linked to relevant commentaries, to articles published elsewhere and to relevant preprints.

Decisions to publish Matters Arising will be taken by journal editors. To ensure the integrity of the published record, and to help readers find all relevant information, published Matters Arising articles will be linked to the online version of the original paper and to the original authors' response. This format will replace Brief Communications Arising as an avenue for post-publication commentary on primary research.

Over the coming months, we plan to introduce the Matters Arising format to the other Nature Research journals, where it will replace Correspondence for such discussions. In this way, we aim to offer a standardized formal mechanism and a constructive peer-review process for post-publication commentary. This should allow debate on published papers in the journals' online pages, and provide visibility and credit for authors engaged in these debates.

All current policies on competing interests, authorship standards (including joint authorship) and author contributions, availability of data, materials and code (where relevant), and publication of the reporting summary will apply to Matters Arising and any published reply from the original authors. ■

Look to Africa to advance artificial intelligence

If AI is to improve lives and reduce inequalities, we must build expertise beyond the present-day centres of innovation, says **Moustapha Cisse**.

Artificial intelligence (AI) is changing society as profoundly as the steam engine and electricity have done. But unlike past technological revolutions, the AI revolution offers a unique chance to improve lives without opening up and exacerbating global inequalities.

That will require widening of the locations where AI is done. The vast majority of experts are in North America, Europe and Asia. Africa, in particular, is barely represented. Such lack of diversity can entrench unintended algorithmic biases and build discrimination into AI products. And that's not the only gap: fewer African AI researchers and engineers means fewer opportunities to use AI to improve the lives of Africans. The research community is also missing out on talented individuals simply because they have not received the right education.

I am happy to be returning to Africa as part of a chance to change that. Next month, Internet-technology company Google will open an AI Research Lab in Accra, Ghana, which will be the first of its kind on the continent and which I will lead. We plan to employ many engineers and scientists.

About 15 years ago, I started undergraduate studies in maths and physics in my native Senegal, and began to teach myself AI using courses downloaded from the Internet. I went to Paris to finish my graduate studies and then took up a position in Facebook's AI research division, working to make machine intelligence fair, transparent and more reliable.

There are many obstacles to an AI researcher from Africa making it into the global community. At a 2016 conference in Barcelona, Spain, attended by more than 5,000 people, I was one of fewer than 10 black people. In response, I co-founded Black in AI. It is now a thriving community of more than 1,000 students, researchers and AI enthusiasts ready to share ideas and foster collaboration to increase representation of black people. Despite the support, many of us still have trouble making it to conferences. I have had papers accepted at meetings but been unable to attend because Western countries such as Australia denied me a visa, even though I was already settled and working professionally in Europe.

We need more efforts to overcome these barriers and to ensure that the benefits of AI arrive globally. Many of the essential ingredients are already in, or coming into place. The human resources are there. Africa is home to the youngest and fastest-growing population on Earth. I am 33 years old, and that makes me older than most of the continent's inhabitants (the median age in Africa is 19; in the European Union, 43). Enthusiasm is huge. Last year, the Deep Learning Indaba gatherings across Africa hosted 300 students from 23 African countries, and had to turn down more applicants than it could accept.

Financial resources are also becoming available. Last year, venture capitalists poured US$560 million into tech start-ups in Africa. Google is supporting and advising more than 60 through Launchpad Accelerator Africa. According to the International Monetary Fund in Washington DC, six of the ten counties with the fastest-growing economies are in Africa.

There is a strong support among AI researchers. Last month, the African Masters of Machine Intelligence degree programme, which I organized with sponsorship from Facebook and Google, started courses with 30 students. More than 30 global experts have agreed to come to the African Institute of Mathematical Sciences in Kigali, Rwanda, to teach it. If the quality of a programme is judged by the quality of the lecturers, this is the best curriculum in the world.

The next step is to develop a coordinated plan to encourage AI education across the continent, incentivize entrepreneurship in the AI sector, and facilitate collaboration between AI researchers and experts in health care, agriculture and other sciences. We need a pan-African strategy: a set of ambitious goals for AI education, research and development and industrialization.

THE OFT-OVERLOOKED CONTINENT HAS MUCH TO GIVE AND TO GET FROM AI.

African nations must forge ties with their local AI expertise. Other countries, including China and France, have benefited from their country-specific initiatives. Canada, in particular, has cultivated a good synergy between a strong AI research community and receptive political leadership. As a consequence, the major tech companies all have centres in Montreal.

Another essential step is setting out a road map to mobilize human and financial resources, including a pan-African AI fund to support coordinated efforts. A network of African institutes of artificial intelligence, for example, could retain the best talents on the continent, enlist world-class African scientists to tackle AI challenges in the African context and collaborate with existing academic institutions. The network could also support policymakers and collaborate with the private sector.

Finally, African governments must create a standard legal framework and a set of values that will help to ensure that AI in Africa serves the good of humanity. There is a growing fear around the world about nefarious uses of AI. Fortunately, initiatives to prevent such uses are emerging. In the public sector, the European Commission this April outlined an approach to setting ethical guidelines for AI. In the private sector, Google published a set of standards this June to govern its AI research and product development.

Now is the time to build a foundation that ensures that AI helps bring better lives in Africa and beyond. With foresight and planning, the technological revolution that AI brings will be a force to empower a fair and prosperous society. The oft-overlooked continent has much to give and to get from AI. ■

Moustapha Cisse *is head and co-founder of the Google AI Research Lab in Accra, Ghana, and professor of machine learning at the African Institute of Mathematical Sciences.*
e-mail: moustaphacisse@google.com

SEVEN DAYS *The news in brief*

POLICY

NIH chimps

Elderly and sick chimpanzees owned by the US National Institutes of Health (NIH) might spend their retirement at research facilities instead of a federal sanctuary if they are too frail to relocate, the agency said on 18 October. The NIH decided to retire all their chimps from research in 2015. Director Francis Collins says that the agency will develop guidelines to determine whether chimpanzees at facilities owned or supported by the NIH are well enough to move. The criteria will include assessments of each animal's health, behavioural, social and environmental needs. If a research facility and the federal sanctuary — called Chimp Haven, in Keithville, Louisiana — cannot agree on whether an animal should be relocated, Collins says, the agency will have a panel of independent veterinarians inform the final decision.

Transparency rule

The US Environmental Protection Agency (EPA) has delayed a decision on whether to implement a controversial rule that would limit the types of scientific research that can be used to justify environmental regulations. The proposed rule, unveiled in April by then-EPA administrator Scott Pruitt, would prevent the agency from basing regulatory decisions on studies whose underlying data are not publicly available — a requirement that could eliminate a lot of epidemiological and other health research. These studies often use confidential patient data that cannot be made public owing to privacy concerns. In a decision quietly released on 17 October, the agency delayed a final determination on the rule until at least 2020. The proposed rule faced criticism from researchers earlier this year, and the agency's scientific advisory board voted to review the proposal in May.

ESA/S.CORVAJA

BepiColombo heads off to Mercury

A joint Japanese–European mission that is set to be the second ever to enter Mercury's orbit launched successfully on 19 October from Kourou, French Guiana. The €1.6-billion (US$1.8-billion) mission, called BepiColombo, will insert two probes into the planet's orbit in 2025, after several fly-bys of Earth, Venus and Mercury itself. One probe, built mainly by the European Space Agency (ESA), will study Mercury's surface and its inner structure. The other, built by the Japan Aerospace Exploration Agency, will focus on the planet's magnetic field and its interaction with the solar wind. Hours after launch, BepiColombo successfully deployed its antennas and two 15-metre-long "solar wings", ESA said. The craft also took several "space selfies" using three on-board monitoring cameras.

Vaccine firm fined

A Chinese pharmaceutical company caught producing faulty rabies vaccines has been fined billions of yuan by national and local agencies. China's national drug regulator announced on 16 October that it will fine Changchun Changsheng Biotechnology 12 million yuan (US$1.7 million), and revoke the company's licence to make rabies vaccines. The regulator determined that when the company created several batches of faulty rabies vaccine, it broke multiple laws, including: using expired products to make the vaccine; not testing the potency of vaccines according to prescribed methods; and destroying evidence to cover up its actions. The drug regulator of Jilin province, where the company's headquarters are based, has issued the firm with a much larger fine: 7.21 billion yuan. The local regulator has revoked the company's licence to produce pharmaceuticals and has banned 14 executives and others involved in the incident from engaging in drug production in the future. The company has not publicly responded to the penalties, but in a statement to its investors, the board said it would set up a group to handle compensation claims.

Brexit warnings

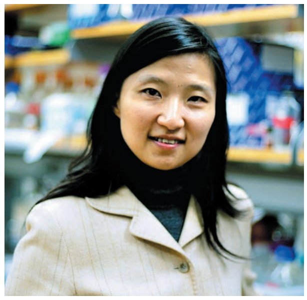
World-leading scientists and mathematicians from across Europe have called on UK and European Union leaders to maintain the "closest possible cooperation" on science after Brexit, and have warned that any barriers to research collaboration in Europe will be to the detriment of all. The calls

were made in a letter signed by 29 Nobel laureates and 6 recipients of the prestigious Fields medal for mathematics, and sent to UK Prime Minister Theresa May and European Commission president Jean-Claude Juncker on 19 October. They come as the United Kingdom approaches its March 2019 deadline for leaving the EU, and amid stalling negotiations on the country's future relationship with the bloc. Brexit could see the end of Britain's participation in EU research programmes — although UK leaders have indicated a desire to be part of future initiatives. The letter says that the United Kingdom must step up its commitment if it wants to stay involved.

AWARDS

Science mega-prize

The inventor of a revolutionary, high-resolution microscopy technique was one of six big winners of this year's Breakthrough prizes, announced on 17 October. The awards, each worth US$3 million, are given out annually in the life sciences, mathematics and fundamental physics. Xiaowei Zhuang (**pictured**), a biophysicist at Harvard University in Cambridge, Massachusetts, received one of the four



CHERYL SENTER/HHMI

life-sciences prizes for leading the development of stochastic optical-reconstruction microscopy — known as STORM — just over a decade ago. The technique was one of the first to break a fundamental resolution limit of conventional light microscopy, and is now used widely by biologists to reveal the hidden molecular structures of cells. Charles Kane and Eugene Mele at the University of Pennsylvania in Philadelphia won the fundamental-physics prize for their work predicting the existence of a type of exotic material known as a topological insulator. See go.nature.com/2nxevs1 for a full list of winners.

PEOPLE

Physicist retires

Lawrence Krauss will retire from his position as a professor at Arizona State University (ASU) in Tempe in May 2019. Earlier this year, the university began investigating allegations of sexual harassment against the physicist, allegations that Krauss denies. He has been on administrative leave since early March, following news reports of alleged harassment. The university accepted Krauss's request to retire, an ASU spokesperson said on 21 October. Krauss founded and led ASU's Origins Project on cosmic questions for nearly a decade; it is now being transformed into an Interplanetary Initiative to focus on the future of humans in space.

HEALTH

Ebola outbreak

The World Health Organization (WHO) declared on 17 October that the Ebola virus outbreak in the northeast of the Democratic Republic of the Congo (DRC) — although extremely worrying — is not a 'public health emergency of international concern.' The organization said that declaring the status would not significantly improve an already robust containment effort. The outbreak began in early August and is centred around the city of Beni. As of 21 October, there have been 203 confirmed cases and 35 probable cases — including 153 deaths. The outbreak is the tenth in the country's history and has been complicated by armed conflicts in the region that are hampering containment efforts. The WHO noted the DRC's long experience in handling Ebola episodes but said that ongoing vigilance is crucial.

Lassa fever

Rats fuelled the largest outbreak of deadly Lassa fever in Nigeria this year, according to the most extensive and rapid genomic analysis of the Lassa virus conducted so far. The study, published on 17 October, eases fears that Lassa had mutated into a super-virus that was spreading swiftly between people (K .J. Siddle *et al. N. Engl. J. Med.* http://doi.org/gfdc4v; 2018). Instead, the viral genomes harvested from 220 patients were surprisingly diverse, indicating that most people had not acquired their infections from someone else. The unprecedented speed of this analysis has helped officials to combat the spread of Lassa fever, and the virus's genetic information will assist researchers as they develop vaccines against the illness. About 514 people fell ill with the disease between January and September, and 134 of them have died. See go.nature.com/2r9vl8x for more.

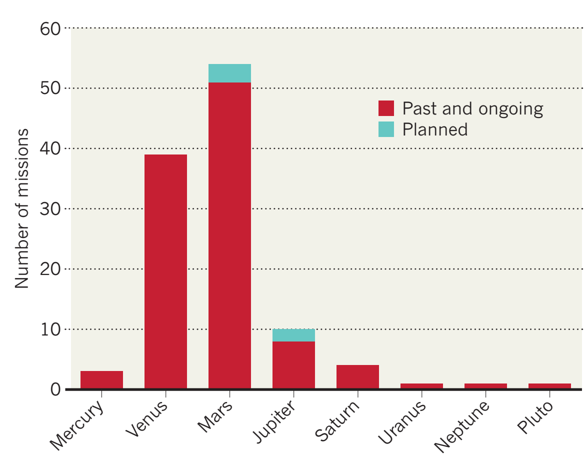
TREND WATCH

The BepiColombo probe, launched on 19 October, will be only the third mission to approach Mercury. Meanwhile, NASA is planning the landing site for its next rover to Mars, a planet that has been targeted around 50 times. So which of our Solar System's planets proved most popular with space scientists — and why? Since planetary missions began in the 1960s, each planet has had at least a flying visit from a spacecraft — but the vast bulk of attention has been focused on just a few. Mars and Venus top the list, partly owing to their proximity and to interest in their past conditions — water on Mars and Venus's previously more Earth-like climate. Mercury has been neglected among the inner planets because it is challenging to reach. Its proximity to the Sun means that craft need to slow down considerably before they enter the planet's orbit. Visiting Uranus, Neptune and the dwarf planet Pluto is also technically challenging, but some astrophysicists say that to truly understand the Solar System, humans will need to step up exploration of these bodies and their moons. Scientists are also planning to send two orbiters to Jupiter's moons.

SOURCE: NASA

PLANETARY VISITS

Nature counts up past, present and future missions to planets and dwarf planets in the Solar System — including failed missions and visits to moons other than our own.



AI ETHICS Survey reveals global differences in moral principles **p.469**

PUBLISHING Chinese government proposes journal 'blacklist' **p.471**

VIROLOGY Scientists race to discover HIV's final hideouts **p.472**

EXPLORATION Researchers prepare to build a Moon base **p.474**

SIMON DAWSON/REUTERS

Could British universities maintain access to EU research funds even with a 'no-deal' Brexit?

RESEARCH COLLABORATION

UK universities go for Brexit gambit to safeguard funds

British institutions set up EU outposts with continental counterparts.

BY QUIRIN SCHIERMEIER

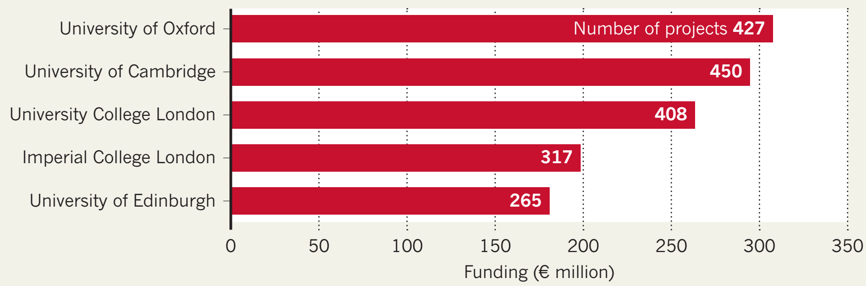
Some of Britain's leading research institutions are establishing alliances with counterparts in other European countries — a move that might allow them to keep drawing on the European Union's science funds even in the case of a 'no-deal' divorce from the bloc, the most extreme form that Brexit could take.

To access the €100 billion (US$115 billion) in research funding that the EU proposes to make available for 2021–27, scientists must be based at host institutions that are legal entities in the EU or associated countries. That might soon cease to be the case for UK universities, many of which stand to lose tens of millions of euros that they get from EU funds at present (see 'Brexit's high stakes'). The outcome depends on the terms of Brexit, which are the focus of intense negotiations.

"In principle, UK entities might remain eligible for funding under EU criteria if they have a legal presence in an EU member country," says Jan Palmowski, director of the Brussels-based Guild of European Research-Intensive Universities. For example, ▶

BREXIT'S HIGH STAKES

Many UK universities stand to lose tens of millions of euros in research funding from EU framework programmes. These are the top five UK universities in terms of income from the Horizon 2020 programme.



SOURCE: EUROPEAN COMMISSION

recipients of grants from the European Research Council must spend at least 50% of their time at a host institute in the EU or associated country. So, continental outposts could help UK researchers to continue to access those grants, even if their country ceases to officially receive them.

Palmowski says that stable alliances with continental partners might also help UK universities to safeguard EU-funded research collaborations and student exchanges. The idea that fruitful research relations built over decades might go to pieces is "dismaying and heartbreaking", says James Conroy, vice-principal for internationalization at the University of Glasgow, which hopes to establish such partnerships.

OXFORD AND BERLIN

Of several alliances launched in recent months, a partnership between the University of Oxford and four institutions in Berlin is so far the most comprehensive. Established at the end of 2017, the Oxford–Berlin Research Partnership is mainly financed by the Berlin state government and private sponsors. This year, the alliance launched a pilot call for proposals and made €10,000–30,000 available in seed grants, with the intention of raising additional third-party funding. Any faculty members of the five institutes can apply. A second call is to be announced next month. Crucially, the partnership will serve as Oxford's legal entity in Germany, and will provide an administrative office at the university clinic Charité in Berlin for visiting researchers. That means, at least in theory, that some Oxford-based researchers might be able to access EU funding. Berlin has also promised to provide space for visiting Oxford scholars in its Natural History Museum.

The likely cost of running the partnership will be around €800,000 a year, says Alastair Buchan, a pro-vice-chancellor and head of Brexit strategy at Oxford and director of the university's Berlin office. And he estimates that this will further enable many millions of euros of research projects and activity. "We're finally doing what we should have done since the day the UK joined the EU in 1973," says Buchan. "We took the freedom to collaborate without restrictions for granted. It was only when the Brexit referendum came along that we began to realize that we must insure against the future."

Oxford and Berlin will both benefit from the partnership, says Steffen Krach, state secretary for higher education and research in the Berlin state government. "Obviously, future access to EU funding for joint research is part of the motivation for Oxford to set up shop here, and quite legitimately so," he says. "But we can also learn a lot from Oxford and their success in scouting international talent. Science in Berlin will doubtless benefit in terms of research output and reputation from lively exchange with one of the best universities in the world."

ACADEMIC ALLIANCES

A host of other similar partnerships are at various stages of development. Institutions involved include the University of Warwick and Northumbria University in Newcastle, as well as the University of Glasgow. Last month, Imperial College London announced an expansion of its long-standing research-and-education partnership with the Technical University of Munich in Germany. "We're naturally interested in any mechanism that allows us to continue fruitful collaborations we have established with European partners over the decades," says Maggie Dallman, vice-president of Imperial College.

EU funding is one way of easing collaboration, but any mechanism to keep doors open in science must be transparent, says Dallman. "We are not seeking to find opaque backdoor routes to getting European funding," she says. "It's ultimately all about doing more research of a higher quality with an outstanding partner."

Conroy says: "Brexit will not leave UK universities unaffected, but we managed to live through turmoil before." He adds: "No matter how difficult the political crisis is, we will see to it that our faculty and students, and society at large, continue to get the best possible scholarship and science." ■

PLANETARY SCIENCE

Mars scientists push for 'mega-mission'

Experts want NASA's next rover to harvest rock at two sites.

BY ALEXANDRA WITZE

NASA's next Mars rover — the first to gather rock samples meant to come back to Earth — should dream big and visit as many places on the red planet as possible, scientists concluded on 18 October.

The rover's stops would probably include some combination of Jezero crater, once home to river deltas and a lake; Northeast Syrtis, which contains some of the most ancient rocks on Mars; and Midway, a compromise option located between the two (see 'Road Trip'). Project scientists have proposed visiting both Jezero, for the river and lake sediments that might retain signs of past life, and Midway, for the ancient rocks. The two are about 28 kilometres apart — so visiting both would be ambitious but achievable.

"The community prefers a mega-mission," says Bethany Ehlmann, a planetary scientist at the California Institute of Technology in Pasadena. "If we're going to do sample return, it has to be a sample cache for the ages."

The Columbia Hills region, which NASA's Spirit rover explored between 2004 and 2011, ranked much lower in the scientists' poll despite having silica deposits similar to those formed by hot springs. "Everybody sort of thought we should go to a new place," says Matthew Golombek, a Mars scientist at NASA's Jet Propulsion Laboratory (JPL) in Pasadena.

The decision about where to send the 2020 rover ultimately rests with NASA's science chief, Thomas Zurbuchen, who will choose in the coming months. "I would be excited about any sample back," says Meenakshi Wadhwa, a planetary scientist at Arizona State University in Tempe. "But we have the luxury of being able to choose between good sites."

Slated to launch in July 2020, the US$2.4-billion rover will be the first from any nation to collect Mars rocks and stash them for a future mission that would bring them back to Earth. The geology of the landing site has to be intriguing enough — and the potential for scientific discoveries there great enough — to make the mission worth the investment.

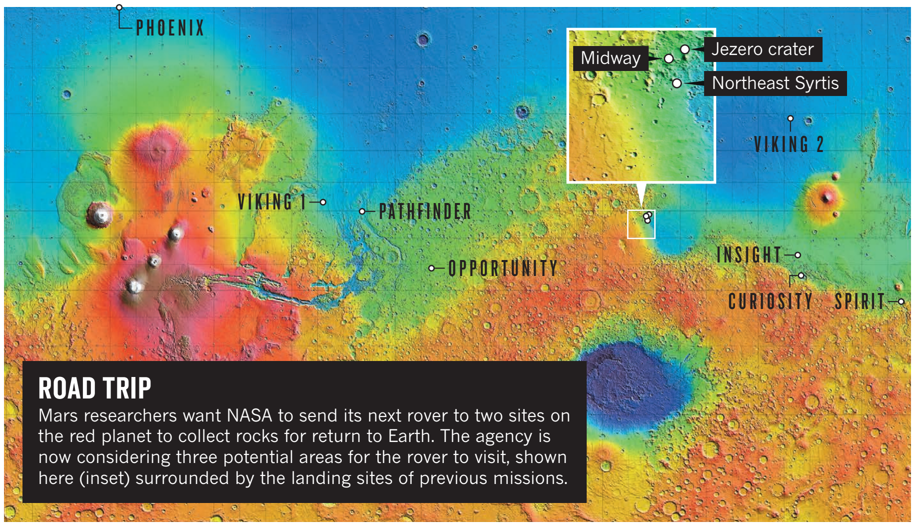
NASA has not planned how it would retrieve the rocks collected by the 2020 rover. But the agency gathered Mars experts in Glendale, California, from 16 to 18 October to hash out the merits of four finalists for the landing site.

Jezero, Northeast Syrtis and Midway came remarkably close to one another in votes by 169 scientists at the workshop. The researchers ranked the sites using several criteria, such as the potential of samples collected at each site to answer crucial scientific questions about Mars.

The idea of visiting Jezero and then Midway — or the other way around — emerged in the past year as mission scientists debated how to get the most out of the rover's journey. "It is ambitious as heck," says John Mustard, a planetary scientist at Brown University in Providence, Rhode Island. Midway's ancient rocks are similar to those at Northeast Syrtis and near the rivers-and-lake system at Jezero.

Sending a rover to Jezero and Midway would mean gambling that the vehicle would last long enough to reach both sites. Its primary mission is 1.25 Mars years (2.35 Earth years); during that time, it is expected to travel roughly 15 kilometres. That would get the rover around most of the Jezero site, if it started there, and possibly even to the crater's rim. But it would then face a trek across dunes to Midway.

NASA's Curiosity rover, the agency's biggest and most powerful so far, has travelled more than 19 kilometres since it landed on Mars in 2012. The engineers developing the 2020 rover expect it to be able to travel faster than Curiosity, in part because of new technology that improves its ability to navigate on its own.



ROAD TRIP

Mars researchers want NASA to send its next rover to two sites on the red planet to collect rocks for return to Earth. The agency is now considering three potential areas for the rover to visit, shown here (inset) surrounded by the landing sites of previous missions.

SOURCE: NASA/JPL

One major question is how many rock samples the rover will collect, and from where. The 2020 rover is equipped with 42 sample tubes, 5 of which will be reserved as spares. That leaves 37 tubes to be filled with the most precious extraterrestrial rocks ever collected.

"Sooner or later, somebody is going to have to decide whether these samples are worth bringing back," project scientist Ken Farley, of JPL, told the meeting. "I don't want to fail because we have not been ambitious enough."

At the workshop, project scientists laid out options for what might fill those 37 tubes. These include chunks of lake deposits from Jezero, fragments of enormous blocks of rock at the crater rim there and samples of the ancient rocks at Midway. The nuclear-powered rover has several possible paths by which to navigate the 28 kilometres of dune fields between Jezero and Midway. Driving that distance would take an estimated 401 Martian days, says deputy project scientist Katie Stack Morgan at JPL.

Still unknown is where the rover might stash its precious samples. One possibility is that it could collect two similar sets of samples at Jezero, depositing one there and carrying the other on to Midway, Farley told the meeting. That would leave open the possibility of retrieving the samples at Jezero if something went wrong with the rover on its way to Midway. Other researchers back a Midway-to-Jezero journey, to get the ancient rocks first.

NASA has not yet decided whether or how it might fetch the samples, although it has tentative plans for a mission in the late 2020s. "We're actually serious about bringing these samples back," Zurbuchen told the meeting. "That's what we're here for." ■

ETHICS

A moral map for AI cars

Survey reveals global variations in ethical rules of the road for autonomous vehicles.

BY AMY MAXMEN

When a driver slams on the brakes to avoid hitting a pedestrian crossing the road illegally, she is making a moral decision that shifts risk from the pedestrian to the people in the car. Self-driving cars might soon have to make such ethical judgements on their own — but settling on a universal moral code for the vehicles could be a thorny task, suggests a survey of 2.3 million people around the world.

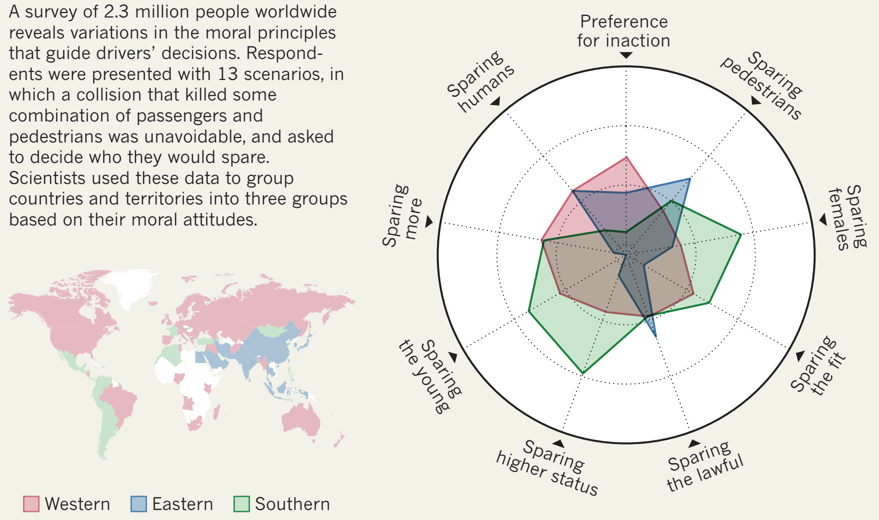
The largest-ever survey of machine ethics[1], published this week in *Nature*, finds that many of the moral principles that guide a driver's decisions vary by country. For example, in a scenario in which some combination of pedestrians and passengers will die in a collision, people from relatively prosperous countries with strong institutions, such as law enforcement, were less likely to spare a pedestrian who stepped into traffic illegally.

"People who think about machine ethics make it sound like you can come up with a perfect set of rules for robots, and what we show here with data is that there are no universal rules," says study co-author Iyad Rahwan, a computer scientist at the Massachusetts Institute of Technology in Cambridge.

The survey, called the Moral Machine, laid out 13 scenarios in which someone's death was inevitable. Respondents were asked to choose who to spare in situations that involved a mix of variables: young or old, rich or poor, more people or fewer.

People rarely encounter such stark moral dilemmas, and some critics ask whether the scenarios posed in the quiz are relevant to the ethical questions surrounding driverless cars. But the study's authors say that the scenarios stand in for the subtle moral decisions that drivers make every day. The findings reveal ▶

MORAL COMPASS



A survey of 2.3 million people worldwide reveals variations in the moral principles that guide drivers' decisions. Respondents were presented with 13 scenarios, in which a collision that killed some combination of passengers and pedestrians was unavoidable, and asked to decide who they would spare. Scientists used these data to group countries and territories into three groups based on their moral attitudes.

ADAPTED FROM REF.1

cultural nuances that governments and makers of self-driving cars must take into account if they want the vehicles to gain public acceptance, they say.

"It's a remarkable paper," says Nicholas Christakis, a social scientist at Yale University in New Haven, Connecticut. The debate about whether ethics are universal or vary between cultures is an old one, he says, and now the "twenty-first-century problem" of how to program self-driving cars has reinvigorated it.

Some of the world's biggest tech companies — including Google, Uber and Tesla — and carmakers now have self-driving-car programmes. Many of these companies argue that the vehicles could improve road safety and ease traffic, but social scientists say the cars raise complex ethical issues.

In 2016, Rahwan's team stumbled on a paradox about self-driving cars[2]: in surveys, people say they want an autonomous vehicle to protect pedestrians, even if it means sacrificing its passengers — but also that they wouldn't buy self-driving vehicles programmed to act in this way.

Curious to see whether the prospect of self-driving cars might raise other ethical conundrums, Rahwan gathered psychologists, anthropologists and economists to create the online Moral Machine quiz. Within 18 months, it had recorded 40 million decisions made by people from 233 countries and territories.

No matter their age, gender or country of residence, most people spared humans over pets, and groups of people over individuals. These responses are in line with rules proposed in what might be the only governmental guidance on self-driving cars: a 2017 report by the German Ethics Commission on Automated and Connected Driving.

But agreement ends there. When the authors analysed answers from people in the 130 countries with at least 100 respondents, they found that the nations could be divided into three groups (see 'Moral compass'). One contains North America and several European and other nations where Christianity has historically been the dominant religion; another includes countries such as Japan, Indonesia and Pakistan, which have strong Confucian or Islamic traditions. A third group consists of Central and South America, as well as France and former French colonies. The first group showed a stronger preference for sacrificing older lives to save younger ones than did the second group, for example.

Test versions of autonomous cars are cruising through several US cities. By 2021, at least five manufacturers hope to have self-driving cars and trucks in wide use.

Bryant Walker Smith, a law professor at the University of South Carolina in Columbia, says that the study is unrealistic because there are few instances in real life in which a vehicle would face a choice between striking two different types of person. "I might as well worry about how automated cars will deal with asteroid strikes," he says.

But Barbara Wege, who heads a group focused on autonomous-vehicle ethics at the car manufacturer Audi in Ingolstadt, Germany, says that such studies are valuable. Wege argues that self-driving cars would cause fewer accidents, proportionally, than human drivers do each year — but that events involving robots might receive more attention.

Surveys such as the Moral Machine can help to prompt public discussions about inevitable accidents, and so might foster trust. "We need to come up with a social consensus," she says, "about which risks we are willing to take." ■

1. Awad, E. *et al. Nature* https://doi.org/10.1038/s41586-018-0637-6 (2018).
2. Bonnefon, J. *et al. Science* **352**, 1573–1576 (2016).

DISASTER MANAGEMENT

Data hint at quake forecasts

Italian-earthquake analysis suggests possibility of predicting aftershocks of some quakes.

BY KATE RAVILIOUS

In 2016, three deadly earthquakes struck Italy between August and late October. Now, an analysis suggests the mechanism that might make such quakes unfold over a period of days or weeks, rather than as a single strike. The conclusion has stirred up excitement among earthquake researchers, because it raises the possibility that seismologists could make life-saving forecasts of the big quakes that follow the first, large quake in a sequence. But challenges remain, including how best to communicate the risk to people who might be affected.

Currently, seismologists can forecast earthquakes only in vague terms — say, estimating a 30% chance of one in a large region in the next 50 years. Most earthquakes take the form of a single large quake followed by aftershocks of decreasing size. But in 'sequence quakes', such as the 2016 Italy event, energy is released in a stop–start manner: several large quakes are interspersed with smaller aftershocks. Scientists aren't sure why this happens.

The latest research — which was published in August and will be presented at the American Geophysical Union Fall Meeting in December in Washington DC — suggests that the answer might lie in the interplay of faults and the movement of underground fluids (R. J. Walters *et al. Earth Planet. Sci. Lett.* **500**, 1–14; 2018). This knowledge could, in theory, be used to predict potentially deadly follow-up quakes.

Sequence quakes occur in all tectonically active areas of the world, but they are thought to be more prevalent in geologically young

fault systems. In Italy's Appennine mountains, which run the length of the country, these quakes occur every few decades, most recently in 2016, 1997 and 1979. More than 300 people died between 24 August and 30 October 2016 as a result of the three earthquakes that hit central Italy, each larger than magnitude 6. The small, historic town of Amatrice was badly damaged by the first quake, and 299 people died.

"Essentially, we can consider sequence quakes as 'failed' big earthquakes," says Richard Walters, a geophysicist at Durham University, UK, who led the research. "The initial stress conditions are the same, but the cascading rupture of multiple segments takes place over days to weeks instead of over seconds."

COMPARE AND CONTRAST

To find out why, Walters and his colleagues took advantage of the wealth of satellite data from the Italian 2016 quakes. The satellites — part of Europe's Sentinel Earth-observing constellation — provided images of the shape of the ground surface. Because the data were taken roughly every 1.5 days, the scientists were able to compare images from before and after each quake and calculate how the ground had moved.

Combining these data with seismological and ground-based measurements, the team found that a network of smaller, cross-cutting faults underlies the Apennine region. The researchers say that these small faults act as barriers to the rupture process, preventing major faults from being 'unzipped' in one go. Had the faults all failed at once, the region would have experienced a single earthquake with a magnitude of about 6.7 — some 50% stronger than the largest individual quake that did strike.

Studying the thousands of small aftershocks that followed the first quake, the team observed tiny quakes creeping northwards at a rate of around 100 metres a day — and found that this matched the speed at which naturally occurring underground fluids would be expected to move along fault lines. "The pattern of small aftershocks suggests that each subsequent quake is triggered by the increased pressure associated with fluids being pumped through the network of minor faults," says Walters. The second quake, two months later, occurred exactly when the aftershocks — and fluid, as predicted by the team's models — reached the next major fault line. "The fluids are being driven by pressure changes. When they reach a fault, the increased pressure 'unclamps' the fault and allows it to move," says Walters.


MAX CAVALLARI/NURPHOTO/GETTY

The Italian town of Amatrice was largely demolished in three earthquakes in 2016.

Nicola D'Agostino, a geoscientist at the National Institute of Geophysics and Volcanology in Rome, finds the mechanism plausible. D'Agostino also agrees that it's theoretically possible to forecast the later quakes, but says it could be tricky to know when a quake is a one-off event and when it is the start of a sequence.

Stephen Hicks, an earthquake scientist at the University of Southampton, UK, thinks that the findings will change how geoscientists work. "Normally, we don't try and interpret aftershocks until later, but I think this will spur scientists into analysing more-subtle features in real time," he says. "The challenge will be to monitor and interpret the data quickly enough to provide a meaningful forecast." ■

PUBLISHING

China's academics await national journal blacklist

But some researchers say the policy won't succeed in improving research quality.

BY DAVID CYRANOSKI

A proposal by the Chinese government to create its own blacklist of journals is creating much debate among the country's scientists, who are still waiting for the list to be revealed, five months after the plan was announced.

Preparation of the list has been shrouded in secrecy. The government says it will include journals that it considers to be of poor quality or those seeking excessive profit, but it has not released its selection criteria, nor has it said when the policy will take effect.

A couple of commercial blacklists exist, and some Chinese institutions already have lists of journals that researchers should avoid, but lists run by government agencies are rare. The Chinese government hopes that a national policy will improve research integrity by reducing the number of low-quality or fraudulent articles from Chinese authors. Academics will receive warnings if they submit to the selected publications.

But some researchers say that a national blacklist won't fix these problems and will be difficult to manage. Lists of approved publications are a better tool for improving research quality, they say.

The science ministry was tasked with creating a blacklist in May, when the government announced a crackdown on scientific misconduct after numerous cases of fake peer reviews, plagiarism and the use of fraudulent data. At the time, the government said that the list would include domestic and international scientific journals, and that publications in these journals would no longer be counted towards a scientist's applications ▶

for promotion, jobs or grant funding.

Nature has seen several lists compiled by Chinese institutions that already tell researchers to avoid certain publications. The Zhongshan Ophthalmic Center at Sun Yat-sen University in Guangzhou circulated a document in January that warned its researchers against publishing in a list of journals that it labelled "controversial in the community" because they have had a lot of retractions. The Obstetrics and Gynecology Hospital of Fudan University in Shanghai has compiled another list. A representative of the hospital says that publications in the journals on its list are not forbidden, but that researchers cannot use grant money to pay the publication fees.

The first list contains two of the world's largest journals, *PLoS ONE* and *Scientific Reports*. Joerg Heber, editor-in-chief of *PLoS ONE*, says that he does not know why some Chinese universities discourage their researchers from submitting to the journal, because only a small fraction (119) of its almost 200,000 papers have been retracted.

A spokesperson from *Scientific Reports* said that they were unable to comment on individual decisions, but hoped that institutions will continue to recognize the journal's value. (*Nature*'s news team is editorially independent of its publisher, Springer Nature, which also publishes *Scientific Reports*.)

Omid Mahian, a thermal engineer at Xi'an Jiaotong University, thinks it would be better to have a national policy that applies to all researchers than for institutions to have their own lists.

And Shi Xiaolei, a science historian at the Institute for the History of Natural Science in Beijing, part of the Chinese Academy of Sciences (CAS), says that a national blacklist and those recently adopted at institutions "will have a positive impact on China's academic environment".

"The real problem is not with the journals, but with the person who submits the article."

Medical researcher Ren Chuanli says that a national blacklist might help to reduce scientific misconduct because it will punish some journals that publish low-quality manuscripts. "But the real problem is not with the journals, but with the person who submits the article," says Ren, who works at the North Jiangsu People's Hospital. He says that some high-quality journals publish low-quality papers, and that some journals that are considered low quality publish the occasional high-quality, highly cited article. Fraudulent papers, too, have appeared in journals of all qualities, so a blacklist based on overall journal quality won't necessarily stop those papers either, says neuroscientist Mu-ming Poo at the CAS Institute of Neuroscience in Shanghai.

Blacklists are also difficult to maintain because new journals are always launching, says Lars Bjørnshauge, the Cophenhagen-based managing director of the Directory of Open Access Journals. Bjørnshauge also wonders whether China might use blacklists as a way of promoting Chinese journals over others.

Yu Liping, who studies academic evaluation at Zhejiang Gongshang University in Hangzhou, doubts that the list will be comprehensive and include all the problematic Chinese journals that have been linked to scientific misconduct. Yu says lists of approved journals that meet certain standards are a better tool than blacklists for improving research standards.

Most researchers agree that what China needs is a more comprehensive system by which to evaluate research quality. "The important thing is whether those who evaluate research are actually evaluating the research and not only looking for papers in 'international' journals on the researchers' CV," says Bjørnshauge. ■

VIROLOGY

Hunting for HIV's hideouts

Scientists are testing new ways of mapping where the virus lurks in the body.

BY SARA REARDON

Antiretroviral drugs have transformed HIV infection from a death sentence to a chronic condition for many people who have the virus. But because HIV never truly leaves the body, the virus rebounds rapidly if patients stop taking the drugs.

Now, scientists are trying to work out how, and where, HIV hides when blood tests show that a person's viral load is low or undetectable. The location of this reservoir has long been a mystery, but that could soon change. Powerful new techniques are giving researchers an unprecedented look at how HIV travels though the bodies of people and other animals — turning up clues to the virus's hiding places and targets for future therapies.

HIV is a challenging foe because it integrates into the DNA of its host cells. Some scientists argue that a true cure would require the removal of all traces of the virus's DNA from the body, rather than simply preventing HIV from hijacking cells to replicate itself — and that goal might be unreachable. "We are starting to realize that getting rid of all the HIV DNA is not completely realistic," says Sara Gianella, an infectious-disease researcher at the University of California, San Diego.

The best that researchers can hope for may be to permanently silence or contain HIV after infection, says Gianella, who last week presented new findings on HIV's hideouts at a meeting organized by the US National Institutes of Health (NIH). Although antiretroviral drug cocktails — known as ART — can suppress HIV in immune cells in the blood, the first results from Gianella's "Last Gift" study confirm that the virus can stash itself in dozens of tissues.

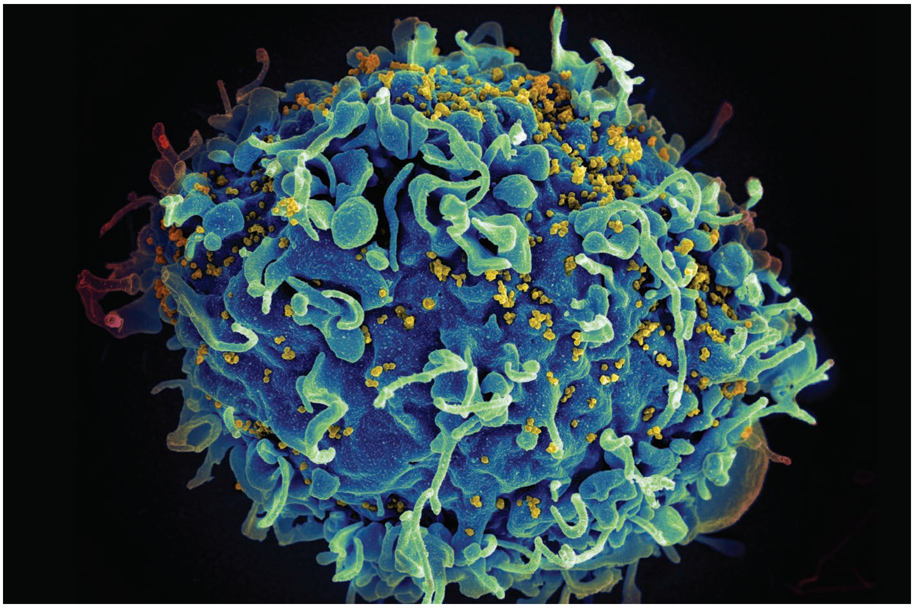
The project examines bodies donated by people with HIV who enrol when they are within six months of death from unrelated conditions. All participants are on ART when they sign up for the study, but some are asked to stop taking the drugs. Gianella's team collects blood samples while donors are alive, and about 50 different types of tissue after death. The samples from people who stopped taking ART show where HIV has rebounded, whereas samples from people who continued on the drugs can provide clues about the virus's reservoir.

The researchers did not detect HIV in the blood of their first donor, who continued taking antiretroviral drugs until his death. But they did find viable virus in nearly all of the 26 tissues they examined after the man died.

Janice Clements, a pathobiologist at Johns Hopkins University in Baltimore, Maryland, calls the project "amazing". After a person with HIV dies, she says, researchers can rarely acquire tissue samples quickly enough to measure the level of virus present, and they usually don't know whether the dead person had been taking antiretroviral drugs.

Clements' research has shown that HIV tends to linger in the brain and cause neurological problems, because most antiretroviral drugs can't cross the blood–brain barrier. At the meeting, Clements and her colleagues presented the first evidence that simian immunodeficiency virus (SIV), which is closely related to HIV, survives in the spinal cord of macaques taking antiretroviral drugs and spreads quickly after the animals stop taking the drugs. "This is unlike any virus we've cured," she says.

NATIONAL INSTITUTE OF ALLERGY AND INFECTIOUS DISEASES/NIH/NCI/SPL



HIV invades immune cells called T cells and uses them to replicate itself.

Nicolas Chomont, a virologist at the University of Montreal in Canada, says that HIV's behaviour in tissues that are known to be part of the reservoir is complex, with virus levels going up and down. "People will tell you the reservoir is everywhere, and that might be true," he says. "Even if it's true, we need to understand if [the virus] in the brain and the big toe are the same."

Tracking these patterns over time might require identifying traces of HIV in the reservoirs of living people — a daunting task. "For these very rare events, you need to know where to look," says Thomas Hope, a cell biologist at Northwestern University in Evanston, Illinois.

At the NIH meeting, Hope presented a new imaging technique in macaques infected with SIV. The researchers inject the animals with antibodies that bind to the virus, which makes it visible in positron-emission tomography (PET) scans of the monkeys' bodies.

The approach has revealed that SIV spreads through mucosal cells in the animals' guts and lymph nodes within hours of infection. Hope has begun to treat infected macaques with ART to determine where and how quickly the drugs lower their levels of SIV. After six months of treatment, the researchers plan to stop the treatment and scan the monkeys to see where the virus has rebounded.

And later this year, another team will begin one of the first PET imaging studies of people with and without HIV in their blood, using a different antibody. Timothy Henrich, an infectious-disease researcher at the University of California, San Francisco, who is directing the study, says that his group hopes to measure what happens when people on ART stop taking the drugs. Gianella's team also wants to test PET imaging in Last Gift participants.

Ultimately, working out where HIV hides will require researchers to do more than just measure the virus in blood, Hope says. Blood immune cells are "going to be a more minor player compared to the dark-matter reservoir", he says. "We know it's there." ■

HOW TO BUILD A MOON BASE

Researchers are ramping up plans for living on the Moon.

BY ELIZABETH GIBNEY

Next year, astronaut Matthias Maurer expects to walk on the surface of the Moon — but without the hassles of a rocket flight, zero-gravity nausea and a risky landing. Instead he'll stroll close to home in a leafy meadow near Cologne, Germany, which is set to host the largest Moon mock-up ever made. On a pit of artificial lunar dust covering more than 1,000 square metres, Maurer and other scientists will be attached to crane-and-pulley systems that allow them to leap as if experiencing the Moon's weaker gravity, and work under adjustable lamps that simulate lighting at different lunar sites. Sometimes, they will retreat to lunar-style living quarters: an airlock-connected module the size of a shipping container.

It's an exciting playground for testing lunar technology, says Maurer, who is a project manager for the multimillion-euro facility. Called LUNA, the mock-up is taking shape outside the European Astronaut Centre in Cologne, with funding from the European Space Agency (ESA) and the German Aerospace Center (DLR). But at 48 years old, Maurer doesn't know whether he will ever put his skills to use on the genuine article. "Hopefully I will make it before retirement. Technically, I believe it's feasible that I will still walk on the Moon," he says.

Maurer's optimism isn't entirely outlandish. He was only two the last time someone visited the Moon for real: US astronaut Eugene Cernan, in the last mission of NASA's Apollo programme. No space agencies have yet committed money to send people back. But, partly as a result of changing political priorities, momentum for human return to the Moon is growing. Rather than rerun the Apollo missions, space agencies are slowly warming to the idea of establishing a sustainable settlement.

Researchers relish the idea of a base for conducting experiments on the Moon and as a way to trial technologies for heading to Mars. Private firms, however, are increasingly tempted by the possibility of mining oxygen and hydrogen — which power rockets — from lunar ice. If that

NASA/GODDARD SPACE FLIGHT CENTER/ ARIZONA STATE UNIVERSITY

does pan out, then the Moon could become a refuelling station, radically reducing the expense of space travel. "Water is the oil of space, and there's mounting evidence that it's there in economically viable deposits," says George Sowers, an aeronautical scientist at the Colorado School of Mines in Golden and previously chief scientist at United Launch Alliance, a firm in Centennial, Colorado, that provides launch services for the US government.

Space agencies are generally reluctant to predict a timeline for a crewed Moon base — in part, because the goal lies well outside their budget horizons, but also because it requires businesses to provide much of the money for the various stages involved. But ESA's director-general, Jan Wörner, has for years talked about multiple countries and companies collaborating in a semi-permanent settlement, which he calls a 'Moon village'. China's National Space Administration has also been quoted in state media saying that a Moon base is its goal, although it has not announced when that might happen.

Lunar exploration prospects gained a boost last December, when a US presidential directive told NASA to switch its sights from exploring asteroids to returning humans to the Moon. NASA has since asked companies to develop lander technology and has outlined plans to request billions of dollars over the next five years for new programmes to support lunar exploration, leading to eventual human return. "It's a fairly radical change of direction," says Sowers. And this October, European engineering firm Airbus launched a contest called the Moon Race, with supporters that include ESA and the US spaceflight firm Blue Origin, which promises to fund companies to develop key technologies for the sustainable development of the Moon.

In the next few years, the only missions to touch down on the Moon will be robotic: China and India will launch probes in the next three months, and Russia has one scheduled within the next five years. But NASA, ESA and the space agencies of Russia, Japan and Canada all support the idea of building a space station in orbit around the Moon by the mid-2020s. (In his 2019 budget request, US President Trump proposed that NASA spend US$2.7 billion on the project over the next 5 years.) That orbiter could provide a base from which to make short, multi-week crewed trips to the Moon's surface in a pressurized rover — which Maurer calls the "camper-van solution". Finally, a settlement could follow. "I think that 20 years is a realistic time frame for a lunar surface infrastructure of some sort, either inhabited by or tended by humans," says James Carpenter, strategy officer for human and robotic exploration at ESA in Noordwijk, the Netherlands.

Researchers have long explored ways to harvest lunar resources, but more in hope than expectation. Now, they are ramping up lunar settlement technology with genuine anticipation that their work might be put into action. While Maurer and others at ESA's LUNA centre practise living on and mining the Moon, others are working on how to grow food and build radiation-proof shelters (see 'Living on the Moon'). At a July meeting at ESA's European Space Research and Technology Centre in Noordwijk to prepare for future human Moon missions, more than 250 specialists from academia, mining, metallurgy, construction and architecture pitched their ideas. "If you ran the same workshop five years ago, it may have only been a handful of people," says Aidan Cowley, science adviser at the European Astronaut Centre. "The appetite has really increased." Although a Moon base might still never happen, Earth-bound preparation for lunar living is well under way.

"WE'D LIKE TO KNOW WHAT IT TAKES TO BUILD LIVING SOILS OUT OF WHAT IS ESSENTIALLY SPACE DUST."

WATER MINING

Lunar settlers' first challenge will be harvesting water. The Apollo missions, which collected samples from the Moon's equator, suggested that the satellite is dry and barren. So the discovery a decade ago that the Moon's poles harbour patches of water ice "was a game changer", says Robert Mueller, a senior technologist at the NASA Kennedy Space Center in Cape Canaveral, Florida, who develops lunar-mining technologies.

For now, researchers don't exactly know where the ice is, how thick it is or whether it is mixed with soil or packed in sheets. India's Chandrayaan-2 rover, scheduled to launch next year, and Russia's LUNA 27 lander, planned for 2022, will target these questions. Russia's lander will have a 2-metre-long drill designed by ESA, and a laboratory to study the origin and abundance of lunar water. NASA also wants to hunt for that water, and has commissioned a suite of companies to develop lunar landers that would carry prospecting instruments, beginning as early as next year. A four-crew human base would need a negligible amount of this water — perhaps dozens of tonnes per year, says Sowers — and there's plenty of it. "Estimates based on current data suggest there may be 10 billion tonnes per pole," he says.

The vast majority of the ice would be mined for fuel. Sowers calculates that mining firms could turn a profit by extracting around 1,000 tonnes of water a year and electrolysing it into its constituent oxygen and hydrogen parts for propellant. The Moon's low gravity means that it would be much cheaper to stock up for long-distance space travel from there than it is from Earth. As an example, a lunar return mission that refuelled at the Moon would cost just one-fiftieth of the price of one that brings all its fuel with it from Earth.

In August, scientists using data from India's Chandrayaan-1 orbiter found that the Moon's ice lies on its surface — but in permanently shaded craters as frigid as −249 °C, the naturally coldest spots known in the Solar System. Excavation machines would need heat and power to release the water and turn it into propellant. Because plutonium-based batteries, which rely on the heat generated by the natural decay of radioisotopes, are too expensive for most private firms, lunar prospectors will probably have to harness energy from the Sun.

They will take inspiration from southern Norway, where giant mirrors set high on a mountain overlooking the town of Rjukan have since 2013 beamed sunlight onto a patch of the town's central square that would otherwise be grey and chilly all winter. Prospectors would hope to do something similar on the Moon, says Sowers. Light from high peaks could be directly channelled into the craters, he says, where it would heat the ice and turn it into vapour. From there, condensed water would be shuttled to a processing plant and split by solar electricity into hydrogen and oxygen. These gases could then be stored and either used as propellant or channelled through fuel cells to supply energy.

Alternatively, rovers could scoop up ice-filled soil and warm it in ovens to release water. The ovens could be powered wirelessly, by training high-power lasers onto photovoltaic cells on the rover. The LUNA facility could test how this would work in real life, with added challenges such as Moon dust in the atmosphere that scatters the laser beams, says ESA's Leopold Summerer. LUNA scientists will also clamber into life-like Moon craters to see how easy it is to navigate these deep, dark slopes, Maurer says.

LIVING OFF THE SOIL

If ice isn't accessible, there is an alternative source of water on the Moon: its soil, also known as regolith. Regolith contains silica and metallic oxides that make it — on average — 43% oxygen by mass, and it is found everywhere on the Moon. Oxygen grabbed from soil could power scientifically or economically interesting outposts far from the poles, and produce useful by-products such as rare metals.

Regolith doesn't give up its riches easily. Releasing oxygen from its chemical bonds is more energy-intensive than heating ice. In theory, a reactor could use giant mirrors to channel sunlight onto a furnace little bigger than an envelope, heating Moon dirt to more than 900 °C until it glows. At that temperature, hydrogen or carbon, brought initially from Earth, can strip oxygen from its minerals and bind the element together

LIVING ON THE MOON

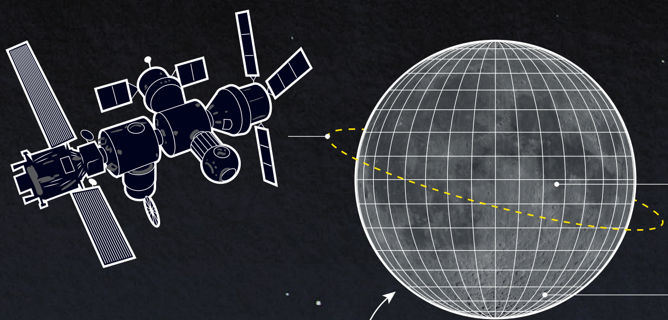
HOW WE MIGHT RETURN TO THE MOON — AND SET UP CAMP THERE.

BY ELIZABETH GIBNEY
ILLUSTRATION BY MACIEJ RĘBISZ
DESIGN BY JASIEK KRZYSZTOFIAK

No one has committed money to send humans back to the Moon. But momentum is building for a lunar return. And scientists are already preparing technologies for a sustainable settlement on Earth's satellite that could generate fuel, water, food and shelter to support a human base and act as a refuelling station for missions farther afield. Here's what it might look like.

LUNAR ORBITAL PLATFORM

Humanity's next international outpost in space might be a lunar orbiter, after the International Space Station is retired, probably in the mid-2020s. Supported by NASA, ESA, JAXA and others, this hub could launch rovers to the Moon and act as a staging post for humans travelling to the lunar surface.



WHERE TO SETTLE

EQUATOR
A base on the equator would be the easiest site to land and launch from, and would be in constant communication with Earth. But lunar nights would prove a challenge for power.

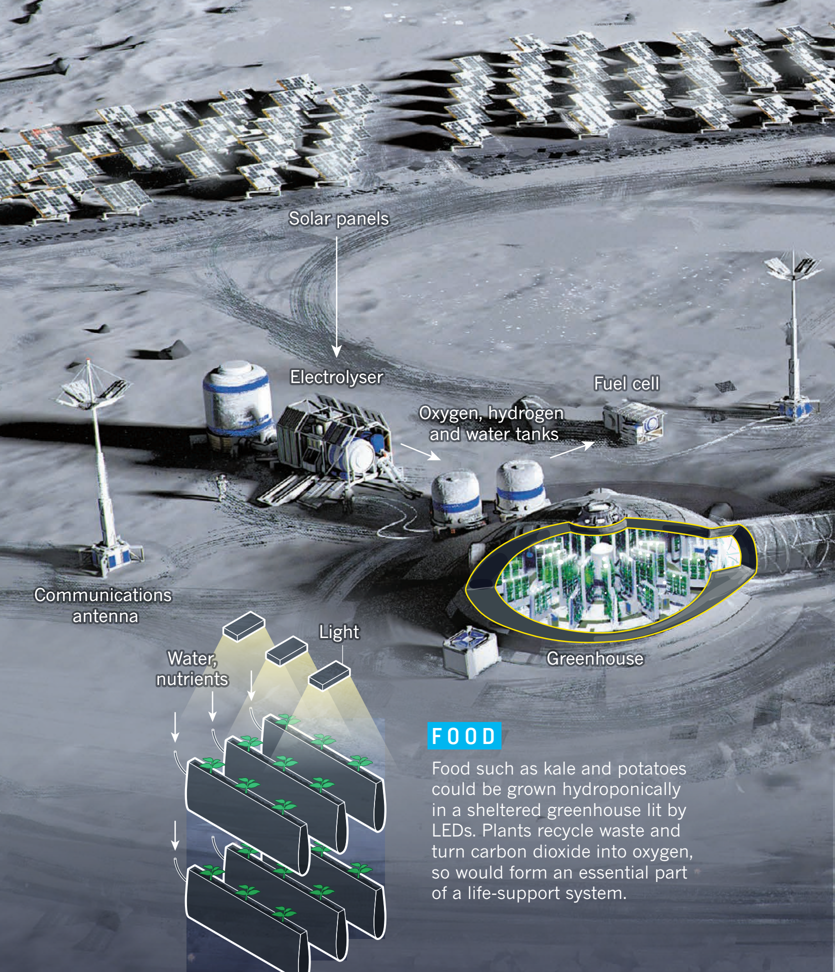
POLES
A settlement in the polar regions offers access to icy deposits for mining, interesting geology and sunlit uplands, but shadowed terrain makes landing difficult, and Earth communications would be intermittent.

FAR-SIDE TELESCOPE

Scientists hope that a telescope might be installed on the far side of the Moon, where sensitive radio receivers would be shielded from interference from Earth.

POWER

Photovoltaic arrays could generate electricity, with solar concentrators providing heat for processes such as 3D printing. Lasers could beam energy from sunlit areas to shadowed regions. Solar-driven electrolysers could split water into oxygen and hydrogen, which can form propellant or be recombined in fuel cells for energy at night.



FOOD

Food such as kale and potatoes could be grown hydroponically in a sheltered greenhouse lit by LEDs. Plants recycle waste and turn carbon dioxide into oxygen, so would form an essential part of a life-support system.

WHEN?

● Scheduled ● Expected ● Speculative

2018–19
China and India launch rover missions to lunar near side and far side.

2020
First commercial lunar landers.

2020
Launch of NASA's heavy-lifting Space Launch System and Orion craft.

2022
Construction starts on outpost in lunar orbit.

2022–23
Russia's LUNA lander heads to a polar region.

2023
NASA's first crewed trip around the Moon in half a century.

2024
First crew inhabits orbiting lunar outpost.

2025
ESA mission to demonstrate water or oxygen harvesting.

Mid–late 2020s
Robotic scouts prepare for human exploration.

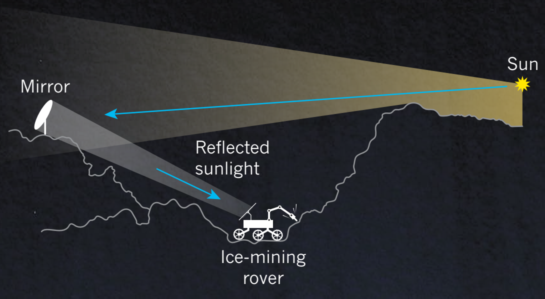
Late 2020s
Short-stay human missions to the Moon in pressurized rover. Tests of lunar mining on large scale.

From 2030s
First habitats with support systems and use of lunar resources.

Late 2030s
Semi-permanent lunar settlement

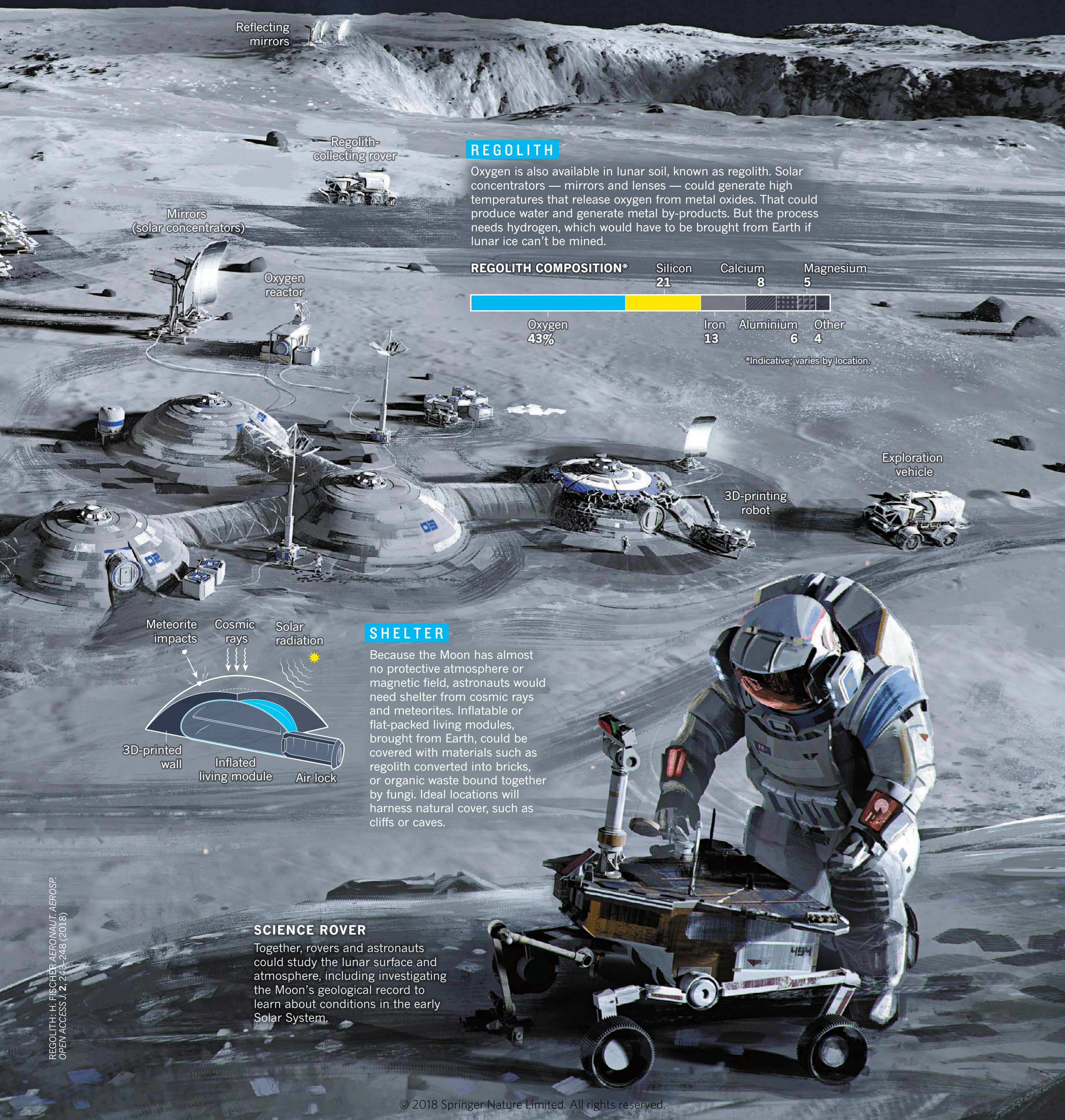
2018 2020 2022 2024 2026 2028 2030 2032 2034

MOON IMAGE ABOVE: NASA



ICE

Surface ice — in permanently shadowed craters — would be the readiest source of oxygen and hydrogen. Sunlight, channelled into the craters with mirrors, could heat the ice, and an overhead dome could capture water vapour. Or rovers could be equipped with ovens that would warm ice-filled soil to make water.



REGOLITH

Oxygen is also available in lunar soil, known as regolith. Solar concentrators — mirrors and lenses — could generate high temperatures that release oxygen from metal oxides. That could produce water and generate metal by-products. But the process needs hydrogen, which would have to be brought from Earth if lunar ice can't be mined.

REGOLITH COMPOSITION*

*Indicative; varies by location.

SHELTER

Because the Moon has almost no protective atmosphere or magnetic field, astronauts would need shelter from cosmic rays and meteorites. Inflatable or flat-packed living modules, brought from Earth, could be covered with materials such as regolith converted into bricks, or organic waste bound together by fungi. Ideal locations will harness natural cover, such as cliffs or caves.

SCIENCE ROVER

Together, rovers and astronauts could study the lunar surface and atmosphere, including investigating the Moon's geological record to learn about conditions in the early Solar System.

REGOLITH: H. FISCHER *AERONAUT AEROSP. OPEN ACCESS J.* **2**, 243–248 (2018)


Prototype rovers are doing test runs in Spain's Canary Islands.

ROBBIE SHONE/ESA

with hydrogen to make water. A field test in Hawaii in 2010 on simulated lunar regolith demonstrated that the process was feasible — although working in low gravity and a vacuum is untested. "Essentially it's a proven technology and would be ready to go within a few years," says Mueller.

Researchers hope to improve the process further to cut down on what needs to be brought from Earth. At the Polytechnic of Milan in Italy, a group led by aerospace engineer Michèle Lavagna is developing a prototype that works at lower temperatures and recycles all the inputs — in this case, methane and hydrogen — so that soil is the only consumable. Currently, one device might take decades to generate enough water to power a single Apollo-style lander back into orbit. But Lavagna says that on the Moon, multiple reactors could work in parallel. With around €600,000 (US$692,000) of ESA funding, her team is now working on a demonstrator plant that would be small enough to fly on a mission.

FINDING SHELTER

If it turns out that water can't be harvested for profit, an outpost based around scientific experiments will still develop, argues Maurer. "Without a commercial perspective, it simply will take much longer to materialize," he says. "We might end up in a situation similar to research in Antarctica — driven mainly by pure scientific interest."

Researchers are excited about the experiments that a return to the Moon could yield, says Robin Canup, a planetary scientist at the Southwest Research Institute in Boulder, Colorado. Sampling the Moon's ancient craters could reveal how the Moon–Earth system formed, when the early Solar System was in a state of flux and asteroids pummelled the satellite. Researchers would also like to study the Moon's water cycle and its seismology — and to install a radio-telescope shielded from Earth interference, which could study radiation from the early Universe.

Unlike in Antarctica, however, lunar residents need to be sheltered from the charged particles of radiation and tiny meteorites that rain down from space — because the Moon has next to no protective atmosphere or magnetic field. The first flat-pack shelters are likely to be brought from Earth, but would need to be covered with metres of sand or regolith, Maurer says. One solution is natural: to exploit cliffs, canyons, caves and lava tubes — tunnels caused by ancient volcanic activity — to protect living quarters. Last year, scientists reanalysing radar data from the Japanese space agency's SELENE orbiter and density measurements from NASA's GRAIL mission found a candidate tunnel that apparently runs for kilometres, beneath the Marius Hills on the Moon's near side. On Earth, researchers have already practised commanding rovers to drive in lava tubes in Lanzarote, in Spain's Canary Islands, to prepare for future lunar exploration.

A few hundred metres from LUNA, in a DLR laboratory in Cologne, Matthias Sperl is testing another idea — growing artificial stone from regolith. In Sperl's lab, a beam of intense light, concentrated on a coin-sized patch, fuses together fizzing sheets of powder at a searing 1,100 °C. Over time, these layers add up to create dark grey, grainy bricks, as in a 3D printing process. On the Moon, sunlight could be concentrated to do the same thing, says Sperl, whose project is part of RegoLight, a €1-million collaboration between the DLR, Belgian aerospace company SAS and architecture and engineering firms. The sheets do not bond together perfectly, but the bricks are already around one-fifth as strong as concrete and comparable to plaster, says Sperl. Architecture firms Bollinger Grohmann Schneider and Liquifer Systems Group, both in Vienna, showed in April that by interlocking the bricks into arches and domes, they could create robust structures. Sperl says these are solid enough to survive Moon quakes and bear the weight of more gravel piled on for protection. Currently, the process takes around 5 hours to make a single brick — but channelling more sunlight could speed it up, he says. Scientists elsewhere are exploring making shelters by searing regolith together in microwave ovens, or binding it together with materials brought from Earth, such as polymers.

KALE DIET

Plant scientists have also spent a good deal of time thinking about the final ingredient needed in a self-supporting Moon base: food. As part of a closed ecosystem, plants would recycle organic waste and turn carbon dioxide into oxygen to breathe. In May, Chinese state media reported that volunteers finished a record 370-day stay inside such an ecosystem, a simulated base known as Lunar Palace 1, in which they grew crops and raised mealworms for protein.

Astronauts on the International Space Station (ISS) already eat space-grown lettuce and other leafy greens. A NASA programme run from the Kennedy Space Center, known as Veggie, has helped to select crops that grow well in confined spaces and are packed with the nutrients that degrade most in storage — vitamins C1, K and potassium. Kale is a winner. "It's a powerhouse that hits everything," says Veggie project manager Trent Smith.

On the Moon, astronauts would grow plants in water under white and red LEDs which they can tweak to alter the mineral and vitamin composition of the plant. Next year, ISS tests will examine how the composition of tomatoes changes depending on the light. More studies will be needed to establish how best to grow crops in regolith's mixture of metals. "We'd like to know what it takes to build living soils out of what is essentially space dust," says Smith. If plants can grow in regolith, adds Veggie researcher Matthew Romeyn, "suddenly that means bringing small fruit trees, not just leafy greens."

"If humans can only stay for short periods of time because of hazards, and if they can't grow food locally, the programme will implode," Mueller says. Another barrier might be a legal one: the 1967 Outer Space Treaty, which all major space-faring nations have signed, states that no country can "appropriate" any part of a celestial body. Most nations today accept that this does not rule out mining in space, says Dimitra Stefoudi, an expert in space law at the University of Leiden in the Netherlands. Since 2015, two countries, the United States and Luxembourg, have enacted national laws allowing space mining, to promote nascent companies. (Russia and Belgium are among countries which argue that mining needs a new international framework.) But the 1967 treaty also says that space activities should be for the benefit of all countries and humankind, so firms will still need to find ways to share know-how and the eventual gains of harvesting resources on the Moon, Stefoudi says.

Ultimately, says Mueller, setting up camp on the Moon is likely to be a desire checked not by technology, but by political will and economics. "If we can solve both of those, I absolutely believe that permanent inhabitation of the Moon will happen." ■

Elizabeth Gibney *is a reporter for* Nature *in London.*

CITIZEN SCIENCE COMES OF AGE

Efforts to engage the public in research are bigger and more diverse than ever. But how much more room is there to grow?

BY AISLING IRWIN

Filip Meysman knew he had made his mark on Antwerp when he overheard commuters discussing his research project on the train. Then, just a few days later, he saw an advertisement about his work on television. There it was, he says, "in between the toothpaste and George Clooney's Nespresso".

As a biogeochemist at the University of Antwerp in Belgium, Meysman wasn't used to drawing so much attention. But that was before he adopted the citizens of northern Belgium as research partners. With the help of the Flemish environmental protection agency and a regional newspaper, Meysman and a team of non-academics attracted more than 50,000 people to CurieuzeNeuzen, an effort to assess the region's air quality (the name is a play on Antwerp dialect for 'nosy' people).

The project ultimately distributed air-pollution samplers to 20,000 participants, who took readings for a month (see 'Street science'). More than 99% of the sensors were returned to Meysman's laboratory for analysis, yielding a bounty of 17,800 data points. They provided Meysman and his colleagues with information about nitrogen dioxide concentrations at 'nose height' — a level of the atmosphere that can't be discerned by satellite and would be prohibitively expensive for scientists to measure on their own. "It has given us a data set which it is not possible to get by other means," says Meysman, who models air quality.

Citizen science — active public involvement in scientific research — is growing bigger, more ambitious and more networked. Beyond monitoring pollution and snapping millions of pictures of flora and fauna, people are building Geiger counters to assess radiation levels, photographing stagnant water to help document the spread of mosquito-borne disease, and taking videos of water flow to calibrate flood models. And an increasing number are donating thinking time to help speed up meta-analyses or assess images in ways that algorithms cannot yet match.

Japanese priest Sadamaru Okano stands beneath a Geiger counter (top left) that sends radiation readings to the Safecast project.

BEHROUZ MEHRI/AFP/GETTY

The movement is surfing wider societal forces, including a thirst for data; the rise of connectedness and low-cost sensor technologies; and a push to improve the transparency and accessibility of science. Increasingly, government institutions and international organizations are getting in on the action. The US and Scottish environmental protection agencies, for example, have incorporated citizen science in their routine work. The United Nations Environment Programme is exploring ways of using citizen science to both monitor the environment and stoke environmental concern. And the European Commission has made a range of funding opportunities available for citizen science within its €80-billion

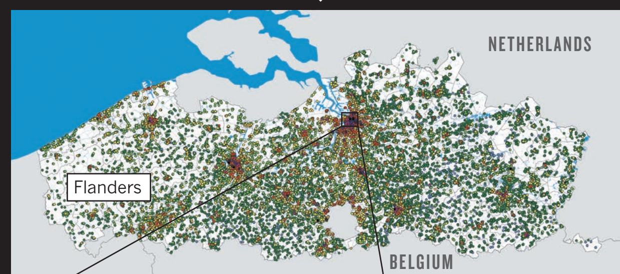
SOURCE: CURIEUZENEUZEN

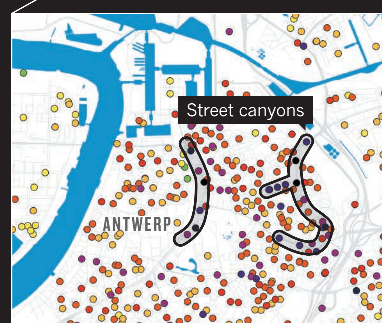
STREET SCIENCE

In May, a collaboration that included the Flanders Environment Agency in Belgium ran a month-long citizen-science campaign to help test a computer model of air quality in the region.

Participants installed nitrogen dioxide samplers on first-floor, street-facing windows inside V-shaped signs to create a standard measurement set-up.

All told, some 20,000 people participated in the project across the Flanders region. Each paid €10 (US$11.5) to join the experiment.





The results are still being formally assessed. But early analysis has revealed — among other things — that the ability of some building arrangements to concentrate traffic exhaust in "street canyons" had been underestimated.

NO_2 concentration

High ○ ● ● ● ● ● Low

(US$92-billion) Horizon 2020 research and innovation programme.

At the same time, citizen-science proponents have grand visions for the future of the field. They hope that such efforts will become a major source of high-quality data and analysis in areas relevant to policy-makers as well as scientists. In December, multiple citizen-science organizations banded together to form a worldwide group — the Citizen Science Global Partnership. One of its first tasks is to explore how citizen science can help to monitor progress towards the UN's Sustainable Development Goals, which aim to address global challenges ranging from hunger to environmental degradation by 2030.

To gain legitimacy, many expect that the field will have to overcome lingering concerns about the reliability of its measurements and its usefulness in research. "There needs to be some type of acceptance and institutionalization of citizen science," says Steffen Fritz, a specialist in Earth observation and citizen science at the International Institute for Applied Systems Analysis (IIASA) in Laxenburg, Austria. "It needs to be not just bottom-up — it needs also to be accepted as some kind of official data stream."

COUNTERS AND ENCOUNTERS

The origins of citizen science go back at least a couple of millennia. In ancient China, migratory locusts frequently destroyed harvests, and residents have helped to track outbreaks for some 2,000 years. The modern form of such research arose after science became a professional activity, creating a cohort of interested outsiders in the process. The phrase 'citizen science' itself was coined in the mid-1990s. Alan Irwin, a sociologist now based at the Copenhagen Business School, defined it both as "science which assists the needs and concerns of citizens" and as "a form of science developed and enacted by the citizens themselves".

Some of the earliest modern citizen-science projects, starting with bird counts in the early twentieth century, involved concentrated outdoor campaigns to record animal sightings. Since then, public involvement has grown to encompass a range of roles. Muki Haklay, a geographer at University College London, has outlined a taxonomy of involvement, from 'crowdsourced' citizen science, in which lay people contribute data or volunteer computing power, to 'co-created' and 'collegial' research, in which members of the public actively engage in most aspects of a project, or even conduct research on their own.

In areas such as biodiversity, where citizen science first thrived, projects are breaking boundaries through the sheer volume of participants and data. The Global Biodiversity Information Facility, the world's largest such repository, says that it gets half of its billions of data points from lay sources. The group estimates that it has supplied data for more than 2,500 peer-reviewed papers in the past ten years.

At iNaturalist, a social network to which anyone can submit a photograph of their encounters with flora and fauna, co-director Scott Loarie has presided over a doubling of submitted images every year since it was launched in 2008. He tries to trace scientists' use of iNaturalist data and has counted 150 papers so far — but he thinks that the actual number is much higher because many of the papers don't cite the organization.

Other researchers have enlisted the public in more-involved projects to enhance research activities, including checking data derived from other sources. When a team published a paper[1] in 2011 suggesting that there could be enough marginal land to grow biofuel sufficient to meet half the world's liquid-fuel needs, Fritz recruited an army of citizen analysers to participate in the IIASA's Geo-Wiki project to study the claim. After working through thousands of images from Google Earth, they generated estimates of land use that were hundreds of millions of hectares lower than those of the original paper[2]. "We downgraded the initial estimates drastically," says Fritz.

Fritz thinks that some people are attracted to his projects because they want to contribute to science, whereas those who become most involved are drawn to the prospect of co-authorship on papers. Some simply like the offer of Amazon vouchers, he says, or a few euros.

Other projects can draw participants for political and social reasons. Within days of Japan's Fukushima Daiichi nuclear disaster in 2011, a small group mobilized to distribute Geiger counters (and ultimately DIY assembly kits) to anyone who wanted to measure radiation levels themselves. At times, local and central governments were hostile to the effort, says Azby Brown, an architect and a leader of the group, now called Safecast. But the findings proved useful, exposing inaccuracies in government readings: high counts where people had been told it was safe to go, and low counts in places that had been deemed unsafe. There is still scepticism about these citizen-generated data, Brown says, although the International Atomic Energy Agency has invited him to speak at several meetings over the past few years.

But it's not just lay people with concerns or scientists with a bright idea who trigger projects: governments and their funding arms are also getting involved. With the support of the European Commission, for example, a project called Ground Truth 2.0 has set up six pilot 'citizen observatories' in Africa and Europe. Each is designed to encourage a three-way conversation between laypeople, scientists (or those who process the data) and those who could benefit from the data, such as policymakers or local authorities. Ground Truth 2.0's leader, Uta Wehn, a researcher at the IHE Delft Institute for Water Education in the Netherlands, says that earlier citizen observatories funded by the European Union included the public as an afterthought. But here, scientists don't dictate the project; they choose the location and let interest groups decide what issue they want to explore and how to do it. "We're putting the people before the sensors," she says.

One observatory, which is examining deteriorating water quality in the Mälaren region of Sweden, found out through early discussions that the existing data on water quality are dispersed, and that local people who do the monitoring had no connection with the decision-makers.

Youth-programme participants Donovan Wooten and Maya Sanders record observations with iNaturalist.

CATIE RAFFERTY/MEDIASANCTUARY.ORG

Two years in, Wehn says it is too early to say whether such projects are changing policy. But participants laud the relationships that have been built between various stakeholders, she says.

Some research leaders are looking to citizen science to foster more inquisitiveness in the 'post-truth' era, in which emotional appeals often seem to win out against fact-based arguments. François Taddei, co-founder of the Center for Research and Interdisciplinarity in Paris, thinks that citizen science can revive critical thinking. Children exposed to such projects are "much less prone to fake news and all these problems that we are facing in the information age", he says.

GROWING PAINS

Yet, even as its aspirations become grander in scale, citizen science faces a number of challenges, including data quality and recruitment — in terms of both persuading more scientists to work on such projects and enlisting enough citizens to participate in them.

Papers published in the past few years have identified flaws in citizen-sourced data, including deviations from standard protocols and biases in recording or in the choice of sampling sites[3,4]. Graham Smith, a wildlife ecologist who analyses sightings made by members of the public for the London-based Mammal Society, a British conservation charity, says that Sunday ramblers will ignore yet another rabbit bounding across their path but unfailingly note a more spectacular sighting such as an otter, which is "the most recorded mammal in Britain for its population size".

Smith, who works for the UK Department for Environment, Food and Rural Affairs, has explored statistical approaches to combat this bias. New apps that track a citizen's route and time in the field are also enriching the data, he says. Meanwhile, simple techniques exist for testing the quality of online analysis, says Fritz. His group inserts occasional control submissions that test a contributor's conclusion against a predetermined professional one (those who regularly fail — about 5%, estimates Fritz — are dropped, whereas those who do well can progress to become co-authors of papers). Scent, a project that uses a gaming app to encourage citizens to photograph land use, has humans and algorithms check one another for errors, says Daniele Miorandi, a communications engineer for the project.

Some academics fear that the public is getting fatigued by all the options, and note that participation in some projects, such as the United Kingdom's long-running Big Garden Birdwatch project, has declined. In an unpublished paper, Haklay has estimated that the number of people globally who could be drawn into regular data collection is about 1.7 million. "You can get a lot of people for a short time investment, or very few people for a deep and intensive engagement, but you can't get everyone doing it all the time," he says.

Researchers and participants are also encountering challenges with ethics, data use and privacy. In Kenya, for example, one of Wehn's citizen observatories is a mapping project that enables people to note poaching incidents, wildlife encounters and fencing, which can be harmful to animals. But the data gathered could be used for nefarious purposes. "Sightings by the tourists might be perfect for the poachers," says Wehn. She says the team is in careful discussion with authorities about what data can be disclosed.

These issues are likely to grow, particularly with the rise of health-monitoring apps. Philip Mirowski, a historian at the University of Notre Dame in Indiana, has raised concerns about the fate of citizen data. He points to projects, such as PatientsLikeMe, that ask people to upload medical information. At least in the United States, he says, "the people who generate the data really don't have any say in what's done with it".

Meanwhile, leaders in the field are pushing for more professionalization, by attempting to systematize the available research and agree on common methodologies. The Open Geospatial Consortium, an international alliance of businesses, research institutes and government groups, has launched a taskforce to get citizen data streams to talk to one another. And the US-based organization SciStarter, an affiliate of Arizona State University in Tempe, has made tools and other resources available for avoiding pitfalls in rolling out projects.

Some are sceptical of efforts to manage citizen science from the top down. Michiel van Oudheusden, a sociologist at the Catholic University of Leuven in Belgium who has studied the example of Fukushima Daiichi, says that citizen science can be especially valuable when it is unaligned with the establishment. "Subversiveness can be very productive," van Oudheusden says.

But Martin Brocklehurst, an environmental consultant and citizen-science advocate, believes that the benefits of bringing order to the field outweigh those of being an outsider. "Too much of citizen science is like a fireworks display: it's great science, but it's short-lived," Brocklehurst says. "We need to start embedding it into the routine way that we do science to support the policy-making process."

Perhaps that is what CurieuzeNeuzen has achieved. The group thinks it reached a world record in the density of air-quality measurements. Now the people of Flanders are mulling over the findings. Among other things, the results revealed that the centres of rural villages, which were thought to have pure air, in fact have high levels of traffic-related air pollution.

The project has opened political doors that more-subdued announcements by the scientific community might never have done. Air quality became a theme in local Flemish elections, which were held in mid-October. Meysman says that he has received many invitations to present his data. And the European Environment Agency says that it aims to apply the approach more widely.

Still, Meysman says, citizen science isn't always feasible. Less-established scientists, under pressure to publish, could not afford the time he has devoted to the CurieuzeNeuzen project, he says. Personally, he has loved watching the effort unfold — the communications campaign, the wave of public interest, the valuable new data — and the chance to put the results to practical and political use. "If I had collected the data myself, I would have had much less impact." ■

Aisling Irwin *is a freelance journalist in Oxfordshire, UK.*

1. Cai, X., Zhang, X. & Wang, D. *Environ. Sci. Technol.* **45**, 334–339 (2011).
2. Fritz, S. *et al. Environ. Sci. Technol.* **47**, 1688–1694 (2013).
3. Tiago, P., Ceia-Hasse, A., Marques, T. A., Capinha, C. & Pereira, H. M. *Sci. Rep.* **7**, 12832 (2017).
4. Kallimanis, A. S., Panitsa, M. & Dimopoulos, P. *Sci. Rep.* **7**, 8873 (2017).

COMMENT

GOVERNANCE Make more use of the patenting system to regulate gene editing **p.486**

ART Pre-Raphaelites interpreted discoveries of a fecund age **p.490**

LAB LIFE Memoir of neuroscientist and equality advocate Ben Barres **p.492**

PUBLISHING Engage more voices in the debate over Europe's open-access plan **p.494**

ILLUSTRATION BY SÉBASTIEN THIBAULT

Do authors comply with mandates for open access?

The first large-scale analysis of compliance with open-access rules reveals that rates vary greatly by funder, report **Vincent Larivière** and **Cassidy R. Sugimoto**.

Last month, European research funders collectively called for research publications to be made free, fully and immediately; so far, 14 funders have signed up. Before that, at least 50 funders and 700 research institutions worldwide had already mandated some form of open access for the work they support. Federally funded agencies and institutions argue that taxpayers should be able to read publicly funded research, and that broader accessibility will allow researchers whose institutions do not subscribe to a particular journal to build on existing research.

However, few empirical analyses have examined whether work supported by funding agencies with such mandates actually is open access[1–4]. Here, we report the first large-scale analysis of compliance, focusing on 12 selected funding agencies. Bibliometric data are fraught with idiosyncrasies (see 'Analysis methods'), but the trends are clear.

Of the more than 1.3 million papers we identified as subject to the selected funders' open-access mandates, we found that some two-thirds were indeed freely available to read. Rates varied greatly, from around 90% for work funded by the US National Institutes of Health (NIH) and UK biomedical funder the Wellcome Trust, to 23% for work supported by the Social Sciences and Humanities Research Council of Canada (see 'Mandates matter').

Our findings have policy implications. They highlight the importance to open access of enforcement, timeliness and infrastructure. And they underline the need to establish sustainable and equitable systems as the financial burdens for science publishing shift from research libraries to authors' research funds.

FREE FOR ALL

Funders with open-access mandates have varying incentives, opt-out mechanisms, copyright protections, deposit guidelines and other associated infrastructures and requirements. These affect when, how and how much work is made open. Our analysis did not assess licensing and instead counted articles found to be freely available to ▶

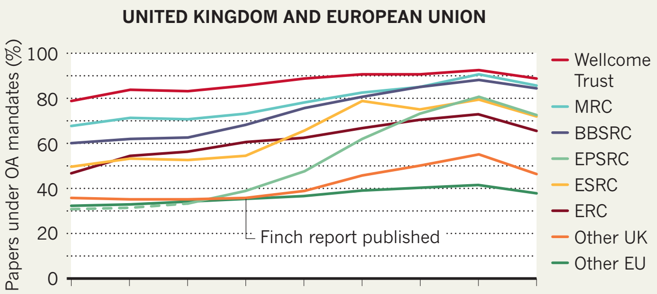
MANDATES MATTER

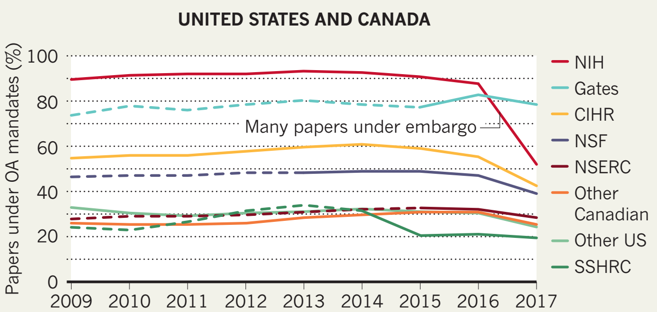
About two-thirds of papers under open-access (OA) mandates are free to read*, either from repositories (green OA) or journal websites (gold OA), with US funders favouring repositories. Of open papers, about half are available by both routes.

VERY VARIED ACCESS

Rates of compliance vary greatly by funder, although they mostly trend upwards. Dips in 2017 are due to embargoes (which delay access for fixed periods after publication).

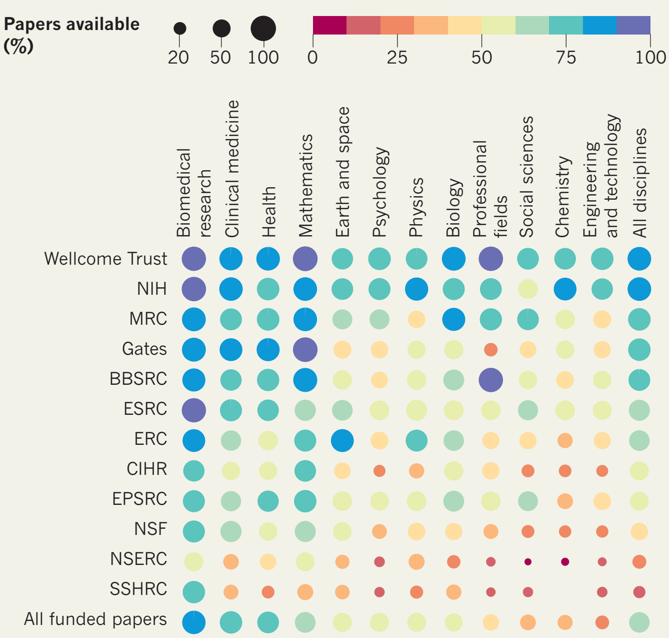
– – – Before mandate adoption





FUNDER EFFECT

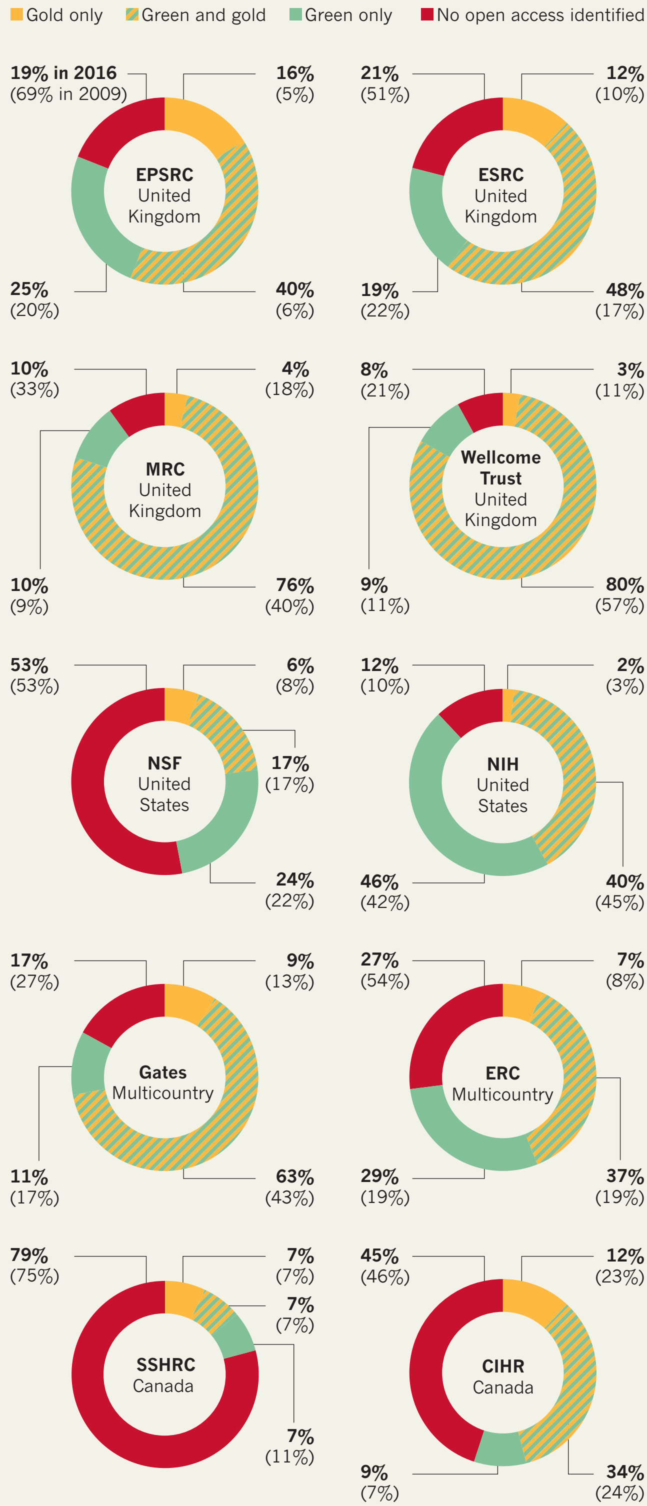
Even within the same discipline, access varies greatly by funder. Of chemistry papers supported by the NIH, 81% were open access; 24% of NSF's chemistry papers were.



MRC, Medical Research Council (UK); BBSRC, Biotechnology and Biological Sciences Research Council (UK); EPSRC, Engineering and Physical Sciences Research Council (UK); ESRC, Economic and Social Research Council (UK); ERC, European Research Council; NIH, US National Institutes of Health; Gates, Bill & Melinda Gates Foundation; NSF, US National Science Foundation; CIHR, Canadian Institutes of Health Research; NSERC, Natural Sciences and Engineering Research Council (Canada); SSHRC, Social Sciences and Humanities Research Council (Canada).

GREEN AND GOLD

Since 2009, the proportion of papers available to read* as both gold and green has soared, even as the proportion of green-only access has stayed relatively constant, and gold-only has dropped.



*Our analysis counted papers freely available to read on publishers' websites as gold and those in repositories as green. It did not consider conditions of reuse or whether free access happened at the same time as publication. Shown over time in Supplementary Information Figure S3.

DATA SOURCES: UNPAYWALL/CLARIVATE ANALYTICS WEB OF SCIENCE/ROARMAP. ANALYSIS: V.L. & C.R.S.

read. We assessed whether access was gold (available on a journal website) or green (available in a repository, such as PubMed Central, the preprint server arXiv or elsewhere, sometimes with a delay or 'embargo' of up to a year after publication). About half (47.5%) of open papers are both green and gold (see 'Green and gold').

Both the NIH and the Wellcome Trust state that they will withhold or suspend payments if articles are not made open access, although it is unclear whether they have done so. These agencies provide convenient repositories for depositing articles: PubMedCentral for NIH-funded work, and Europe PubMed Central in the case of Wellcome. Their policies encourage compliance and allow authors to publish in journals that do not permit articles to be available immediately without a subscription. Although articles must be in a repository at the time of publication, free access can occur later. For example, a paper with a 12-month embargo published in the March 2016 issue of a journal would become freely available in the repository in March 2017.

Funders that allow authors to deposit papers after publication see lower rates of compliance, presumably because authors lose track of this obligation. The Canadian Institutes of Health Research (CIHR) mandated deposit on publication from 2008 to 2015, but dropped this requirement when the three main Canadian research councils adopted a joint, harmonized policy. Compliance for CIHR-funded studies went from around 60% in 2014 to around 40% in 2017.

Other funders that have lower rates of compliance than the NIH and the Wellcome Trust provide less enforcement and infrastructure. For example, the US National Science Foundation (NSF) called for "voluntary compliance" with open-access mandates until early 2016 — its rate was around 47% in 2016. Its repository, which uses infrastructure developed by the US Department of Energy, has less visibility and fewer articles than PubMedCentral does. However, that might soon change, because deposition in this public repository is now mandatory for papers arising from NSF funding awarded after January 2016. Compliance at the CIHR is hampered by similar barriers. Unlike PubMed Central in the United States, PubMed Central Canada was never the dedicated infrastructure for Canadian medical papers. The Canadian repository faltered and then closed in February this year, and no strong environment of enforcement has arisen. Factors include lower funding in Canada compared to the United States, which makes it harder for authors to allocate funds for article-processing fees.

The United Kingdom has seen a steep rise in open-access compliance across all agencies (see 'Very varied access'). Rates at all four of the UK research councils studied went up by at least 20 percentage points between 2009 and 2016; the Engineering and Physical Sciences Research Council went up by 50 percentage points. This follows the publication of the Finch report in 2012 (see go.nature.com/2yojrkc) by a working group of academics, funders and publishers that was established in 2011 by David Willetts, then the UK science minister. It strongly recommended increasing access to research through article-processing charges and gold open access rather than by archiving papers in repositories. For the next assessment of research institutions in England in 2021, major UK funders have now decided to consider only open-access publications.

ANALYSIS METHODS

How we mined data on open-access compliance

We first identified the funding sources of papers using the published acknowledgements (mandated by most funders). These have been indexed by the Web of Science (WoS) since 2008 for science and medicine, and since 2015 for social-sciences articles. There is no uniform format, so we looked for variations of agency names (such as 'NSF' and 'National Science Foundation') and aggregated these.

Next, we used Unpaywall, a platform that helps researchers to find open-access articles. It identifies the population of scholarly papers using the list of unique digital object identifiers (DOIs) registered by Crossref, a non-profit indexing organization. Unpaywall mines all journal websites listed in the Directory of Open Access Journals, along with databases such as PubMed Central and 50,000 other journal websites and repositories. It intentionally excludes papers that are available on social-networking sites (such as ResearchGate) or illegally (such as on Sci-Hub). As of April 2018, Unpaywall provided the open-access status of nearly 96 million scholarly documents.

Of the 12,495,074 journal articles we matched with Unpaywall using DOIs, 1,352,918 acknowledged funding from 1 of the 12 funders we identified.

To determine rates of compliance, we matched Unpaywall data to our set of WoS articles and analysed them by funder and discipline. WoS includes papers published in about 12,500 journals annually, so some funded work is in journals not covered by our analysis, especially in the social sciences and humanities. Our ability to assign funders to papers is imperfect, given the various ways in which funder names appear and because authors do not always provide funding information. Rates of estimated compliance are likely to be conservative; there might be funded papers that are freely available online but which could not be found by Unpaywall. **V.L. & C.R.S.**

FIELD CULTURE

We find variations by discipline, with nearly full compliance in biomedicine, clinical medicine and health research. The social sciences, chemistry and engineering all show lower rates (see 'Funder effect'). Within the same discipline, compliance varies drastically by funding agency. For example, in chemistry research, 81% of work funded by the NIH is publicly available, whereas that is true of only around one-quarter of chemistry studies supported by the NSF and CIHR. Different funders support different types of work, but the variations we found also remain consistent within sub-disciplines (see Supplementary Information, Figure S5). Although researchers cite norms and needs within disciplines as a reason not to comply with open-access mandates, we believe that the funding agency is a stronger driver of open access than is the culture of any particular discipline.

NEXT STEPS

If funding agencies have their own data on compliance, the information should be openly published so that it can be used in assessments of the march of open access, such as ours. That would also allow comparisons to be drawn. Future research on compliance with open-access mandates should evaluate the utility of other data sources, such as Scopus, 1findr, Kopernio and Dimensions (run by Digital Science, a firm operated by the Holtzbrinck Publishing Group, which has a share in *Nature*'s publisher. *Nature* is editorially independent of its publisher). We must also create stronger reporting systems so that these data are more readily available for analysis. This involves collaboration between funders, publishers and indexers. Reporting should allow for analyses at the level of funded projects, which would provide information on the time between funding and open dissemination. On a broader level, more research is needed to understand what makes scientists comply with funder mandates and why.

Ultimately, open access needs a sustainable financing model. Libraries and other organizations have historically borne the cost of publishing through subscription fees. Gold open access displaces those costs on to authors (who often need to allocate funds from their research budgets to cover publishing), even as libraries continue to shell out for subscription fees. The cost of publishing in open-access journals ranges from less than US$100 to more than $5,000 per article, with dominant publishers such as Elsevier averaging $2,612 per paper in

article-processing charges and Springer Nature (which publishes *Nature*) averaging $1,913 (see go.nature.com/2cn3zuy). The system as a whole risks charging multiple actors for the same product, and could price some places and people out of publishing.

Advocacy must be balanced with evidence in the open-access debate. Our research demonstrates that funders can clearly shape compliance through their mandates, and that this compliance needs to be monitored. Real barriers — such as infrastructure and funding — must be overcome to make mandates efficient. However, the rhetoric surrounding disciplinary barriers might be more a myth than a reality: when the proper structure and incentives are in place, researchers comply.

To move the conversation forward, we need a greater sense of the implications of open access on the scientific system's financial structure. We must study how certain publishing models will put pressure on some parts of the system while alleviating it from other areas, or even enriching them. We need to ensure that the mandates are sensitive to financial inequity across countries, disciplines, institutions and researchers.

Universities, industry and funding agencies should think collectively about robust and scalable models. Cooperation and foresight are the only ways to ensure that everyone has open access to research — both for readers who want to consume it, and for authors who wish to publish it. ■

Vincent Larivière *is associate professor at the University of Montreal and associate scientific director of the Observatory of Science and Technology, Montreal, Canada.* **Cassidy R. Sugimoto** *is associate professor of informatics at Indiana University Bloomington, USA.*
e-mail: sugimoto@indiana.edu

1. *Nature* **508**, 161 (2014).
2. De Groote, S. L., Shultz, M. & Smalheiser, N. R. *PLoS ONE* **10**, e0139951 (2015).
3. Gargouri, Y., Larivière, V., Gingras, Y., Carr, L. & Harnad, S. Preprint at https://arxiv.org/abs/1206.3664 (2012).
4. Vincent-Lamarre, P., Boivin, J., Gargouri, Y., Larivière, V. & Harnad, S. *J. Assoc. Inform. Sci. Technol.* **67**, 2815–2828 (2016).

Supplementary Information accompanies this article: see go.nature.com/2yitfgn. C.R.S. did this work while at the US National Science Foundation, funded by an independent research programme.

IMAGINECHINA/REX/SHUTTERSTOCK

A gene-edited 'micropig' was developed in 2015 by the BGI genomics institute in Shenzhen, China.

Use the patent system to regulate gene editing

Governments should use patents to shape the deployment of CRISPR–Cas9 as they have done for past technologies, argues **Shobita Parthasarathy**.

Next month, researchers, policymakers, ethicists and social scientists will meet in Hong Kong for the second International Summit on Human Gene Editing.

Since the first summit, held in Washington DC nearly three years ago, researchers have continued to apply the versatile gene-editing technology CRISPR–Cas9 to diverse domains — from crop enhancement and pest eradication to human disease. Many have flagged the ethical, economic and environmental concerns raised by manipulating plant and animal genomes, including our own. But, so far, governments have struggled to develop viable approaches to regulation.

A crucial part of the arsenal for shaping

the future of gene editing is hiding in plain sight: the patent system.

In the past, patents have played an important part in regulating new technologies and research, from the atom bomb to work involving human embryonic stem cells. Some organizations and individual researchers using CRISPR–Cas9 are already creating licensing agreements that reflect their own moral codes. In my view, government-driven efforts centred on national patent systems should be deployed to help regulate gene editing.

NEW LAWS NEEDED

Last year, the US National Academies of Science, Engineering, and Medicine recommended that clinical trials involving gene editing in human eggs, sperm or embryos should be permitted only for the treatment and prevention of serious disease or disability. They also urged that a "stringent oversight" system be developed to limit the use of the technology in this context[1]. In July, the Nuffield Council on Bioethics, a highly respected bioethics body in the United Kingdom, similarly stated that the use of heritable genome editing "could be ethically acceptable" only after appropriate governance measures are put in place[2].

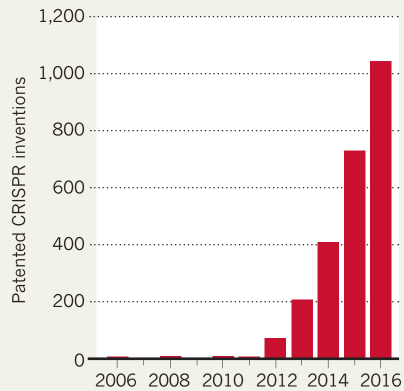
These recommendations haven't yet translated into legal frameworks or formal governance structures. And the history of regulating emerging biotechnologies suggests that such laws could be a long time coming, if they end up being formed at all[3]. For now, when it comes to editing genes in humans and other organisms, the United States and the United Kingdom — along with many other countries — rely on laws and policies that cover existing genetic-engineering technologies. Or, as in the case of human germline editing in the United States, the government simply bans the use of federal funds for such research.

Such policies have been criticized for decades as being inadequate[4]. Their insufficiencies are considerably more problematic in the context of gene editing, which, largely thanks to the development and uptake of CRISPR–Cas9 (see 'Invention protection'), promises to have much greater societal impact than previous technologies for modifying genomes.

In the United States, for instance, the oversight provided by the 'coordinated framework' (developed in the 1980s to deal with genetically engineered organisms) handles only immediate risks. The framework covers the management of altered plants and animals that have already been created, and does not consider the socio-economic, ecological or ethical consequences of creating organisms not found in nature. Likewise, since 2016, a condition in the budget for the US Food and Drug Administration has prohibited the agency from authorizing clinical trials in which a "human embryo is intentionally created or modified to include a heritable genetic modification". But there is nothing to stop US researchers using private funds to edit the genes of human embryos in the lab.

INVENTION PROTECTION

The number of patents for inventions involving CRISPR–Cas9 has soared in just a few years.



Data for 2017 and 2018 are not yet complete; patent applications are generally published 18 months after filing at a patent office.

SOURCES: IPSTUDIES, SWITZERLAND; FAMPAT-QUESTEL

HISTORICAL PRECEDENTS

How could patents help? These legal instruments — which give inventors the right to prevent others from commercializing their technologies — are usually seen solely as contracts that incentivize innovation. In fact, they can do much more, directly and indirectly.

They can lead to higher prices for products, for instance, and reduce people's access to important technologies if inventors use them to establish and maintain monopolies. Perhaps most importantly, they can shape innovation trajectories. Patent laws were a major factor in the 'war of the currents' in the 1880s, driving people to favour engineer George Westinghouse's alternating current (AC) over the direct-current system invented by Thomas Edison. (Westinghouse licensed the US patents for AC from inventor Nikola Tesla.) The decisions that governments make about whether to grant patents implicitly demonstrate their moral approval of an invention and indicate what types of technology are likely to generate exclusive markets[5].

> *"Patents can shape innovation trajectories."*

The idea that governments could use patent systems to shape both the development of a technology and its impact on society is not new. In the 1940s, the US Congress used the patent system to control the development and commercialization of atomic weaponry.

To try to reduce the possibility of private actors developing atomic bombs, or of US intelligence leaking, Congress created a three-tier system of non-patentable, government patentable and privately patentable technologies in the Atomic Energy Act of 1954 (ref. 6). The US Patent and Trademark Office offered standard patents for technologies that fell into the 'privately patentable' category. But inventions that would be useful only in the production of fissionable material, or when using such material or atomic energy in a military weapon, were non-patentable. The government (specifically, the Atomic Energy Commission) could also step in and require 'compulsory licenses' for technologies deemed to be in the public interest.

Even further back, in the nineteenth century, the governments of several European countries, including France, Switzerland and Italy, limited or even banned patents on foods and pharmaceuticals to ensure that people had sufficient access to these products[7].

EXISTING FRAMEWORKS

Biotechnology, including gene editing, is already regulated to some degree through patents.

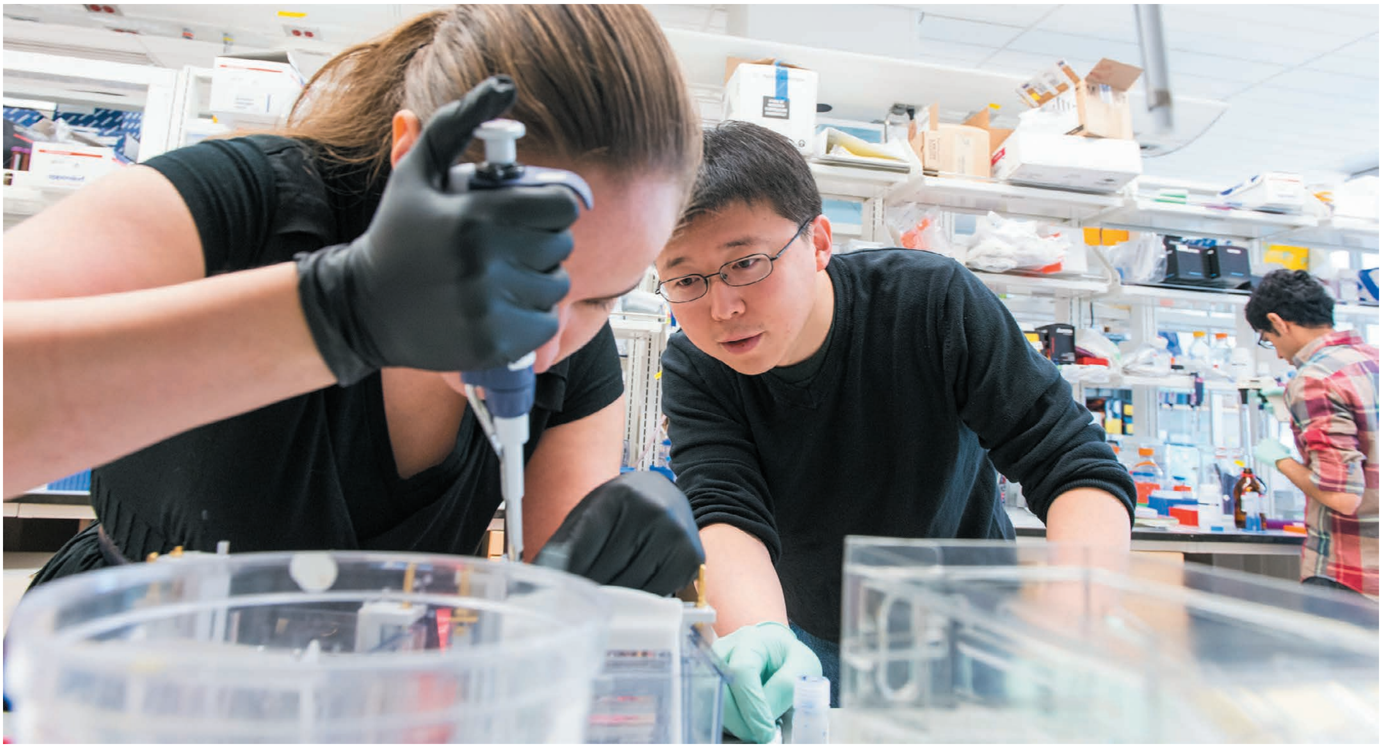
In 1998, the European Parliament and Council passed a directive on the legal protection of biotechnological inventions. This harmonized Europe's approach to patents in the emerging field; it covers all European Union countries and the 38 member countries of the European Patent Office. It also addressed people's concerns about the moral and socio-economic implications of individuals being able to obtain patents on living entities, such as human embryos or genetically engineered plants and animals.

The directive states that governments can grant patents on animals that have been modified only if the resulting benefit to humankind outweighs the animal's suffering. It even includes prohibitions on patenting processes that could be used to modify human sperm, eggs or embryos.

Moreover, some scientific organizations and researchers who use CRISPR–Cas9 have themselves recognized the power of patents to govern gene editing, and are writing their own licensing agreements.

For example, the Broad Institute of MIT and Harvard, in Cambridge, Massachusetts, is a non-profit research institution that holds expansive patents on CRISPR–Cas9 technology. It prohibits its licensees from using CRISPR–Cas9 to modify human embryos, alter ecosystems or modify tobacco plants[8]. Similarly, Kevin Esvelt at the Massachusetts Institute of Technology (MIT), also in Cambridge, holds a patent on a 'gene drive' that could be used to spread a particular genomic alteration throughout an animal population. He requires those who wish to license this patent to disclose their proposed use, and has suggested that other scientists working on gene drives do the

JUSTIN KNIGHT/MCGOVERN INST.



Silvana Konermann (left) and Feng Zhang helped to develop CRISPR–Cas9 at the Broad Institute in Massachusetts, which holds many CRISPR patents.

same. He argues that this will enable public discussion[9].

What I'm calling for, however, is different: more-formal, comprehensive, government-driven regulation using the patent system.

This would cover all domains of gene editing, not just certain areas of research. It would have more transparency and political legitimacy than individual efforts ever could, by involving government institutions that are explicitly charged with representing the public interest. And it would enable governments to exploit the unique vantage point that patent offices have on the early stages of scientific fields and industries. (Inventors usually file patent applications before they try to get regulatory approval for new technologies.)

In the United States, Congress could authorize a working group to convene an advisory committee for gene-editing patents. The working group could include: individuals from the Environmental Protection Agency, who are trained in assessing ecosystem impacts; staff from the Department of Commerce, which oversees the US Patent and Trademark Office; personnel from the Department of Health and Human Services, who have deep understanding of biomedical research, health-care costs and research ethics; and staff from the Government Accountability Office, which in the past few years has developed expertise in technology assessment. The advisory committee should also comprise scientists, physicians, ethicists, social scientists, historians, lawyers and representatives from the private sector.

Building on existing laws such as the 1954 Atomic Energy Act, the committee could put together a regulatory framework for reviewing and awarding patents related to gene editing. It would need to incorporate the perspectives of citizens at every step[10], and might place inventions into distinct categories. Perhaps the use of CRISPR–Cas9 for editing human embryos would not receive patent protection, for instance, whereas the use of the technology to correct a common mutation that causes heart failure would.

Under such a framework, the committee could identify inventions that are likely to be so important to the public interest that the government should monitor closely how associated patents are used and licensed, and step in to force broad licensing if a patent holder charges too high a price for access to their invention. (Currently, the 1980 Bayh–Dole Act gives the US government 'march-in' rights in the case of taxpayer-funded research, although it has never been used in this way[11].)

The EU directive on the legal protection of biotechnological inventions already provides Europe with some guidance on which gene-editing processes and products to exclude from patentability[5]. But 20 years on, additional oversight is needed. To develop a more detailed governance framework, the European Patent Office should convene an advisory committee to develop a framework, similar to the one proposed for the United States. This could then by adopted by the European Patent Office and EU member countries.

Ultimately, patent law will need to be just one of many regulatory schemes. Some developers might still create and use ethically problematic technology, even if they are unable to patent it. But existing approaches, and the entities that are conventionally tasked with overseeing areas of scientific research, seem ill-equipped to address complex societal and value-based concerns in an increasingly privatized world. Patents, which affect the thousands of investigators now using CRISPR–Cas9 in both the private and public sector, should be part of the mix. ■

Shobita Parthasarathy *is professor and director of the Science, Technology, and Public Policy Program at the Gerald R. Ford School of Public Policy, University of Michigan, Ann Arbor, Michigan, USA.*
e-mail: shobita@umich.edu

1. National Academies of Sciences, Engineering, and Medicine. *Human Genome Editing: Science, Ethics, and Governance* (National Academies Press, 2017).
2. Nuffield Council on Bioethics. *Genome Editing and Human Reproduction: Social and Ethical Issues* (Nuffield Council on Bioethics, 2018).
3. Kelly, S. E. *Sci. Technol. Hum. Val.* **28**, 339–364 (2003).
4. Jasanoff, S. *Designs on Nature: Science and Democracy in Europe and the United States* (Princeton Univ. Press, 2005).
5. Parthasarathy, S. *Patent Politics: Life Forms, Markets, and the Public Interest in the United States and Europe* (Univ. Chicago Press, 2017).
6. Riesenfeld, S. A. *Calif. Law Rev.* **46**, 40–68 (1958).
7. Cassier, M. *Hist. Technol.* **24**, 135–151 (2008).
8. Guerrini, C. J., Curnutte, M. A., Sherkow, J. S. & Scott, C. T. *Nature Biotechnol.* **35**, 22–24 (2017).
9. Regalado, A. 'Stop "Gene Spills" Before They Happen' *MIT Technology Review* (20 October 2016).
10. Jasanoff, S. & Hurlbut, J. B. *Nature* **555**, 435–437 (2018).
11. Eberle, M. *Marquette Intel. Prop. Rev.* **3**, 155 (1999).

To paint *Ophelia* (1851–52), John Millais spent months observing plants on the banks of the Hogsmill River in Surrey, UK.

HISTORY

Rebels of art and science

John Holmes explores the empirical underlay to Pre-Raphaelite masterworks.

Extraordinary advances in science, technology and industry shaped the Victorian age; alongside that grew a new experimentalism in literature and the arts. From 1848, the Pre-Raphaelite Brotherhood, a group of British artists founded by Dante Gabriel Rossetti, William Holman Hunt and John Everett Millais, began to weave science into their art. They sought a new aesthetic even as they called for art to model itself on science — and were championed by scientific luminaries from the comparative anatomist Richard Owen to physician Henry Acland.

The Pre-Raphaelites rejected the insistence of the Royal Academy of Arts in London that artists should learn by imitating the paintings of Raphael. Modern interest in the group has grown steadily since a revival among the counter-culture of the 1960s. That is newly reflected in a retrospective on the work of Edward Burne-Jones (who, with Rossetti and William Morris, formed the 'second-wave' Pre-Raphaelite movement) at Tate Britain in London. Burne-Jones's paintings, such as the 1880s Briar Rose series featured in the exhibition, seem to open a window on an exquisitely romanticized fantasy world caught in moments of stillness. But John Ruskin, the era's leading art critic and a serious amateur geologist and botanist, saw something else there. In 1884, he wrote that, although the brotherhood's work might "seem to be the reaction of a desperate fancy … against the incisive scepticism of recent science", they were in fact "a part of that science itself". Ruskin supported the Pre-Raphaelites and knew them well.

Edward Burne-Jones
Tate Britain, London. Until 24 February 2019.

From the start, they insisted on working "from absolute data of fact" and acute observation, as the critic William Michael Rossetti — brother of Dante Gabriel and poet Christina Rossetti — explained in the magazine *The Spectator* in 1851. He noted that the group conducted "investigations" through art, and offered the public "unflinching avowal of the result". A year previously, Pre-Raphaelite art critic Frederic George Stephens had spelt out the group's commitment to these principles in its short-lived periodical *The Germ*. Stephens remarked that since the early 1800s, disciplines such as chemistry had made astonishing progress "by bringing greater knowledge to bear upon a wider range of experiment", and pursuing precision. Why, he asked, shouldn't the same methods benefit the arts' "moral purposes"?

How did the group harness empirical methods to create its work? Take arguably the most famous Pre-Raphaelite painting, Millais's *Ophelia* (1851–25). At first glance, this seems a sentimental portrayal of the tragic suicide of the character in William Shakespeare's *Hamlet*. However, every plant depicted, from purple loosestrife to wild roses, is the product of more than three months of painstaking observation as Millais worked on the banks of the Hogsmill River in Surrey. Other artists had painted in the open air before, but never in such meticulously wrought detail.

This became a collective experiment to discover what painting, pushed to its limits, could reveal. Each new work would press further, recording exact effects of light and shade, as in Hunt's 1851 *The Hireling Shepherd*; or ecological relationships and

animal behaviour, as in the straying sheep of his 1852 *Our English Coasts*; or skin tones in full sunlight, as in Ford Madox Brown's *The Pretty Baa-Lambs* (1851). Subjects were scrupulously researched. Hunt visited Jerusalem and the Dead Sea to study the landscape, people and latest archaeological findings for his paintings of the life of Jesus.

Looking more closely at *Ophelia*, we see a study of physical and psychological phenomena. Millais asked his model, Elizabeth Siddall — poet, artist, and later Dante Gabriel Rossetti's wife — to lie in a bath fully dressed. Candles were set under the bath for warmth. Millais's concentration was such in his epic eight-hour bout of painting that he failed to notice when they went out; Siddall caught a severe chill and he paid the doctor's bill. Siddall thus paid a price for Millais' 'laboratory conditions', but the method did enable him to capture how hair and fabric float on and underneath the surface of water.

Ophelia's expression in the painting is also revealing. The Pre-Raphaelites were rightly scathing about the state of psychology in 1850, when phrenology and physiognomy still passed as sciences. Stephens called it "dry operose quackery ... mere chaff not studied from nature, and therefore worthless, never felt, and therefore useless". They set out to study the mind through art instead, refining their designs as they thought through the mental states of their subjects. An early drawing for *Ophelia*, now in the Plymouth City Museum, is melodramatic. In the finished painting, we see a much more subtle analysis: Ophelia subdued by despair, sinking into unconsciousness as she drowns. When Stephens argued that scientific methods could advance art's moral purpose, this is what he meant: it helps us to understand humanity and nature.

The art establishment was crushing in its opposition to the group, as were mainstream journalists. Charles Dickens called Millais's provocative 1849–50 painting *Christ in the House of His Parents* "odious, repulsive, and revolting". By contrast, many Victorian scientists supported the brotherhood. The naturalist William Broderip, who bought *The Hireling Shepherd*, introduced Hunt to Owen, founder of the Natural History Museum in London (and coiner of the word 'dinosaur'). Owen became a staunch advocate of the Pre-Raphaelites, and delighted in showing Millais and his children around the British Museum's natural-history collections. In 1881, Hunt painted the magnificent portrait of Owen now in the Natural History Museum.

Acland, one of Owen's students and from 1858 the Regius Professor of Medicine at the University of Oxford, was even more central to the movement. When he and Ruskin campaigned in the 1850s for a natural-history museum in Oxford, Acland declared that it would be decorated on Pre-Raphaelite principles. The stonework was based on real plants and animals, carved in minute detail by unsung heroes of Victorian sculpture John and James O'Shea. The stone columns around the central court were geological samples drawn from quarries around Britain to illustrate different periods of Earth's history. The court was surrounded by statues of scientists at work, from Galileo Galilei to James Watt. Rossetti advised on the project; Ruskin and Siddall, among others, contributed designs. Thomas Woolner, the brotherhood's only sculptor, with the group's close collaborators John Lucas Tupper and Alexander Munro, fashioned the sculptures. The Oxford University Museum of Natural History stands as one of the best and most surprising collections of Victorian public sculpture, and the only one dedicated to science.

"Many Victorian scientists supported the brotherhood."

Woolner went on to collaborate with the architect Alfred Waterhouse. When Waterhouse built the commanding Natural History Museum in London for Owen in the 1870s, they again applied Pre-Raphaelite principles. Owen supplied Waterhouse with specimens from the vivarium and illustrations of extinct animals as models for the terracotta menagerie that adorns the museum's facade. Waterhouse paid tribute to Owen by including ancient fauna he had described, such as the archaeopteryx and the palaeotherium.

It was also Woolner who made the Pre-Raphaelites' most direct contribution to science. When carving a bust of Charles Darwin (now in the herbarium at the University of Cambridge), Woolner alerted the biologist to a small protuberance sometimes visible inside the rim of the human ear, known as the auricular tubercle. In the 1840s, when devising a statuette of Puck from Shakespeare's *A Midsummer Night's Dream*, Woolner had observed that the feature also appears in monkeys, as pointed ears. Both Darwin and Woolner recognized this physiological phenomenon as evidence of human evolution from earlier primates, and Darwin — dubbing it the 'Woolnerian tip' — mentioned it in *The Descent of Man* (1871).

Woolner finished the bust in 1869. The same year, the astronomer Norman Lockyer founded a new periodical — *Nature*. Lockyer had Pre-Raphaelite connections, too: he had worked with William Rossetti on an earlier journal, *The Reader*; befriended Hunt; and employed Pre-Raphaelite landscape painter John Brett to accompany an expedition to Sicily to study the solar eclipse in 1870. In 1878, Lockyer wrote a series of articles on 'Physical Science for Artists', setting out guidance in optics and assessing paintings from the latest Royal Academy exhibition on grounds of scientific accuracy. In *Nature*, Lockyer held artists to exacting scientific standards, just as Stephens had done in *The Germ*.

The Pre-Raphaelites launched the most radical and ultimately the most influential Victorian art movement, inspiring the European symbolists and the Arts and Crafts movement led by Burne-Jones's great associate, William Morris. They also took their lead from — and shaped the culture of — Victorian science, and affected its legacy to this day. ■

OUMNH

Ferns carved by James O'Shea top a column at the Oxford University Museum of Natural History.

John Holmes *is professor of Victorian literature and culture at the University of Birmingham, UK. His latest book is* The Pre-Raphaelites and Science.
e-mail: j.holmes.1@bham.ac.uk

MEMOIR

Ben Barres: gender champion

Marc Freeman lauds the neurobiologist's posthumously published memoir.

An unstoppable force of nature, unfazed by headwinds, managing to will all of us onwards and upwards: this was Ben Barres. A highly influential neurobiologist and advocate for women in science, Barres lived an unusually interesting life. He was an openly transgender faculty member at Stanford University School of Medicine in California, and a pioneer in understanding the functions of glia — the most abundant and mysterious cells in the brain. Whether by design or accident, along the way he also became a hero for people from gender and sexual minorities (LGBT+ people), and for early-career scientists generally.

In 2017, Barres died of cancer at the age of 63. His posthumously published memoir, *The Autobiography of a Transgender Scientist*, documents his remarkable life story.

Born Barbara Barres in mid-1950s New Jersey, Barres was a precocious, science-loving child who relentlessly pursued academic opportunities. But from an early age, he never felt comfortable being treated as female. As he writes: "internally I felt strongly that I was a boy. This was evident in everything about my behavior." Although educated and trained at a time when sexism was rampant in schools and academia, his talent, self-confidence and drive propelled him swiftly upwards. He entered top-notch medical and science training programmes, from undergraduate study at the Massachusetts Institute of Technology in Cambridge to a medical degree at Dartmouth College in Hanover, New Hampshire, a doctorate at Harvard University in Cambridge and a postdoctoral position at University College London (see A. D. Huberman **553**, 282; 2018).

Barres admits that during his academic career as Barbara, especially early on, he did not even consider the idea that gender would be used to assess anyone's qualifications or limit their opportunities. When a professor accused Barres of enlisting a boyfriend to complete a difficult problem (because he assumed a woman couldn't have done it), Barres was incensed — but mainly about the accusation of cheating. After he transitioned in 1997 and began to live as a man, he came to understand this and other episodes as personal experiences of pervasive sexism. An advocate was born.

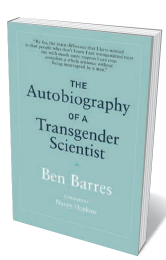
Barres' promotion of women in science is well known, and he recounts some of his efforts in the book. In an essay in *Nature*,

Neuroscientist Ben Barres died last year.

TIMOTHY ARCHIBALD

'Does gender matter?' (B. Barres *Nature* **442**, 133–136; 2006), he helped to debunk the idea of intrinsic gender differences in scientific ability. Barres worked relentlessly to improve the representation of women in all areas of science: as faculty members and conference or departmental seminar presenters, and in leadership positions. To support that aim, he volunteered for countless selection committees, editorial boards and grant-reviewing panels, and spoke up. He would be combative if needed — and that was fun to watch. The effect he had was impressive and inspiring.

More recently, Barres became a vocal advocate for trainees. He describes in glowing terms how grateful he was to his own mentors — David Corey and Linda Chun at Harvard, and Martin Raff at University College London — and the powerful impact of their approach to science on his professional growth. He describes his joy in mentoring young scientists, and the thrill of watching them succeed; he saw mentoring as a sacred duty. His reflections on the importance of picking a good mentor and how we should treat postdoctoral fellows should be required reading for all academics. An appendix lists his trainees, an impressive group including many leaders in their fields. These people were Barres' family.



The Autobiography of a Transgender Scientist
BEN BARRES
MIT Press (2018)

Barres was best known for his work on glial cells, and he describes his lab's major scientific contributions, including his surprising discovery that these cells release factors that help make synaptic connections between neurons. Just as in conversation, Barres' enthusiasm for the science and the people doing it leaps out, and the details of many key discoveries come fast and furiously. Barres was a leader in boosting the status of glia from boring support cells for neurons to essential nervous-system cells that interact with neurons dynamically to help modulate their growth, connectivity, function, plasticity and health. His lab helped to launch a now-thriving field.

It is touching to read Barres' account of the emotional pain he experienced because of his gender discordance. He grappled with this issue at a time when — just as some still do today — certain sectors of society struggled to talk about homosexuality, never mind issues of gender identity. Before he transitioned, he writes, there were moments when he considered killing himself, a situation that is alarmingly common among people with gender dysphoria. He felt he was faced with the choice of his career or his personal happiness.

Science is hard enough to begin with; I can't imagine also having these sorts of life pressures. Fortunately, he had wonderfully supportive colleagues and friends across the scientific community. Because of his own past pain, Barres felt he owed it to the LGBT+ community to discuss his life openly and emphasize how transitioning had brought him relief and happiness.

Barres is as open in memoir as he was in life, and the book teaches important, deceptively simple lessons. Be yourself; be happy; don't apologize for who you are. Be respectful, but be honest and express your opinion even (or especially) if it's not popular. Science is exhilarating, and we have a responsibility to do it well, with fairness to all involved. *The Autobiography of a Transgender Scientist* shows the way. ■

Marc Freeman *is a neuroscientist and director of the Vollum Institute at Oregon Health and Science University in Portland.*
e-mail: freemmar@ohsu.edu

Correspondence

Global responsibility for publishing costs

As readers, many scientists in Europe will welcome the news that most work will have to be published in open-access journals from 2020 (M. Schiltz *Front. Neurosci.* **12**, 656 (2018); see also *Nature* **561**, 17–18; 2018). But as knowledge producers, I fear that many more scientists around the globe are likely to be disenfranchised by richer nations, institutions and funding bodies.

Open-access publication requires authors to pay in the region of US$1,000–3,000 (more than the cost of many research projects in some disciplines). Although scientists from low-income countries are eligible for full-fee waivers, compulsory open access will force many others to use money intended for research, or to publish in low-tier journals that still retain reader paywalls.

In my view, sources of all publication fees should be recorded — just as funding sources are now — so that marginalized researchers can be identified and rates of waiver use tracked. The findings would guide realistic fee capping by European open-access publications.

John Measey *Stellenbosch University, Stellenbosch, South Africa.*
john@measey.com

Plan S debate is not "a pity"

Robert-Jan Smits declares it a "pity" that arguments about academic freedom are stifling debate on his 'Plan S', which promotes a radical shift towards open-access publishing (see *Nature* **562**, 174; 2018). In fact, the opposite is happening.

Spirited debates on the topic are ongoing among researchers, publishers, librarians, journalists, funders and members of the public (see, for example, go.nature.com/2qtusrb; go.nature.com/2coxgrx; go.nature.com/2nm2dmq; go.nature.com/2ckhnrc; go.nature.com/2qw2hv6). We have yet to reach agreement on what to make of the major European funders' radical shift to compulsory open-access publishing by 2020, but we continue to explore this important issue in good faith.

In a Plan S world, the research community will need to address academic responsibility, the future of scholarly societies and their journals, and how to respect disciplinary differences and ensure the high quality of publications. We invite Smits and all other architects of the plan to engage academics in constructive discourse on these issues.

J. Britt Holbrook *New Jersey Institute of Technology, Newark, New Jersey, USA.*
Stephen Curry *Imperial College, London, UK.*
Shina C. L. Kamerlin *Uppsala University, Uppsala, Sweden*
holbrook@njit.edu

Bullying: report without reprisal

I appreciate efforts by institutions to counteract academic bullying (see *Nature* **560**, 420 (2018); *Nature* **560**, 529; 2018). They should also set up a clear, fair and accessible reporting system, with no fear of reprisal for the institution or the people who have been abused.

Bullying issues are arguably worse for international students and scholars than for domestic lab members. International researchers are already disadvantaged by visa requirements and financial constraints, and such abuse exacerbates their insecurities over position and job prospects — particularly if it takes the form of infringement of intellectual property and unfair authorship positioning on publications.

An efficient reporting system for victims would also benefit their institutions and funding organizations by helping them to select a new generation of more nurturing leaders. The media could also play a part by making it clear in their reporting that individual cases are not an indictment on an institution's overall reputation. This might also speed up the handling of abuse cases.

Morteza Mahmoudi *Harvard Medical School, Boston, Massachusetts, USA.*
mmahmoudi@bwh.harvard.edu

Co-producers: move into charity sector

The James Lind Alliance has a 14-year track record of involving patients, carers and clinicians in determining priorities for health research (see www.jla.nihr.ac.uk). Charity Futures, another co-produced research initiative, is using the James Lind Alliance's consultation process for the first time outside medicine.

We ask charities and donors about the research topics that they consider the most important (see go.nature.com/2pwsre4). Our aim is to encourage more research into those areas and so enable charities and donors to base their work on better evidence (see also C. Fiennes *Nature* **546**, 187; 2018). We shall report publicly on our findings next year.

Stephen Bubb *Charity Futures, London, UK*
Katherine Cowan *Brighton, UK*
Caroline Fiennes *Giving Evidence, London, UK*
caroline.fiennes@giving-evidence.com
Competing interests declared (see go.nature.com/2ph6tz3 for details).

Co-producers: frame only the questions

I work at the interface of science and public policy, so appreciate the importance of public values in prioritizing research problems (*Nature* **562**, 7; 2018). The challenge is to make this happen without disrupting the evidence base that enables effective delivery of solutions.

There is a general principle here that traces back to philosopher David Hume's famous distinction between 'is' and 'ought', 'fact' and 'values' (*A Treatise of Human Nature*, 1739). In a democracy, it makes sense that the way we frame what is studied and how we respond to results is subject to dialogue that affords more importance to those affected than to those doing the research. However, values should frame the questions, not the answers.

This is less clear-cut in the social world (where co-production has its roots) than in the natural world, because in the social world observed and observer are not so distinct from one another.

The answers need to be grounded in the scientific treatment of carefully collected evidence.

Peter Calow *University of Minnesota, Minneapolis, USA.*
pcalow@umn.edu

Let writers author scientific literature

Footnotes listing individual author contributions to research papers help to offset ambiguities in formal authorship, but are easily overlooked. Until due credit can be fairly allocated by artificial-intelligence algorithms (see G. L. Kiser *Nature* **561**, 435; 2018), I propose confining authorship to those who wrote the paper. People assessing credit for all other functions would then be forced to consult the detailed contributions list.

This 'authorship for authors' scheme would promote scientists' writing skills. A journal article is a short story: it needs creativity, clarity, structure and pace. Yet scientists receive little training in narrative and tend to write poorly. This impairs progress because readers are forced to struggle with tedious or confusing text. Let's help the scientific literature live up to its name.

Tobias I. Baskin *University of Massachusetts, Amherst, USA.*
baskin@umass.edu

SOLAR PHYSICS

A window onto the Sun's core

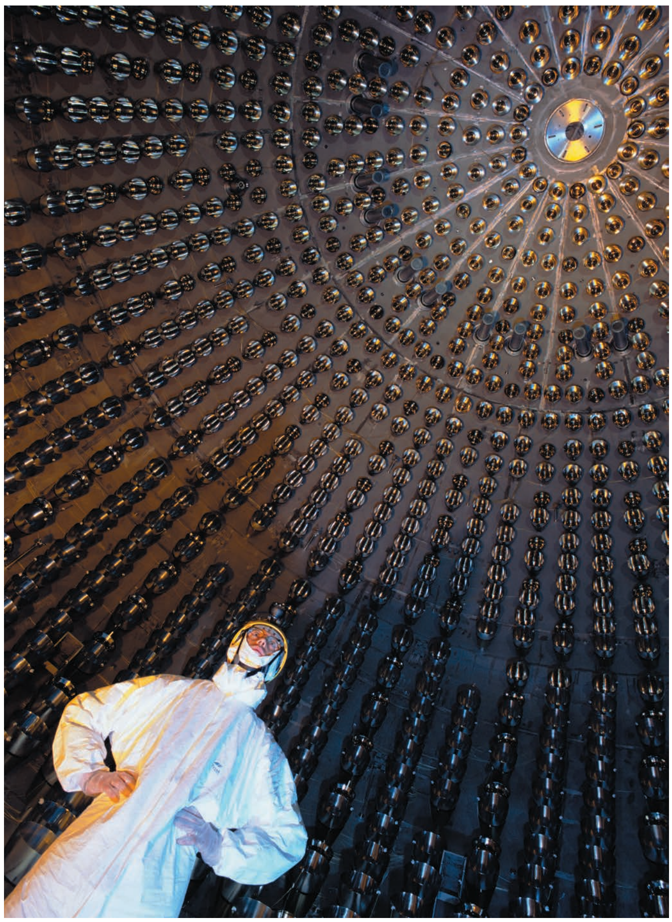
An experiment has measured the energy spectrum of solar neutrinos associated with 99% of the nuclear reactions that power the Sun. The results provide a glimpse into the depths of the solar core. SEE ARTICLE P.505

ALDO SERENELLI

Energy is generated in the interior of the Sun through sequences of nuclear reactions in which four protons fuse together to form a helium-4 nucleus. These sequences are accompanied by the release of two particles known as electron neutrinos. Models suggest that 99% of the nuclear energy released by the Sun originates from three reaction sequences — collectively known as the proton–proton (*pp*) chain — that are initiated by the fusion of two protons. On page 505, the Borexino Collaboration[1] reports the first complete measurement of neutrino fluxes that originate from these three sequences, based on an analysis of more than 2,000 days of data collection. The results help us to understand the details of how and why the Sun shines.

Neutrinos interact weakly with matter, and therefore escape almost unhindered from the Sun's interior, to reach Earth about eight minutes later. Solar neutrinos therefore provide a direct view into the nuclear furnace in the Sun's core. The Borexino experiment (Fig. 1) detects such neutrinos and determines how much energy they have by measuring the amount of light produced when the particles interact with the detecting agent (an organic liquid, called the scintillator, which is kept underground to minimize the amount of background radiation that can interfere with the neutrino signals). In contrast to all other solar-neutrino experiments, Borexino can measure the energies of both high- and low-energy neutrinos, which makes it possible to study the structure of the solar core using a technique known as neutrino spectroscopy.

Electron neutrinos can change into two other types (or flavours) of neutrino, known as tau and muon neutrinos, as they travel to Earth, a phenomenon known as flavour oscillation. The Borexino experiment is more sensitive to electron neutrinos than to tau or muon neutrinos, and so flavour oscillation needs to be accounted for when the measured neutrino fluxes are used to calculate the fluxes produced in the Sun. Taking this into consideration, the Borexino collaborators used the measured neutrino flux to work out the total power generated by nuclear reactions in the Sun's core, with an uncertainty of about 10%, and found that this is the same as the measured photon output — thus showing that nuclear fusion is indeed the source of energy in the Sun. This value, calculated for the amount of energy produced through nuclear reactions, is comparable with previous[2] results obtained by combining data from several neutrino-detection experiments, and places the most robust and model-independent constraints on the source of solar energy.



VOLKER STEGER/SPL

Figure 1 | The Borexino experiment, Gran Sasso, Italy. A researcher stands in a spherical vessel that forms part of the Borexino neutrino detector. The Borexino Collaboration[1] has used the detector to produce the first simultaneous measurement of the neutrino fluxes associated with the nuclear reactions that account for 99% of the Sun's energy.

The findings also have interesting ramifications for neutrino physics. By combining their data with predictions from standard solar models, the collaborators determine a quantity known as the electron neutrino survival probability (which describes the probability that an electron neutrino created in the Sun will also be detected as an electron neutrino at the detector) for neutrinos produced in four reactions of the *pp* chain. The calculated survival probabilities include the best available value for low-energy neutrinos, which correspond to an energy regime in which flavour oscillation is expected to occur mostly in vacuum conditions. Combined with the survival probabilities determined for higher-energy neutrinos, the findings give strong support to our current understanding[3,4] of neutrino oscillations — that is, the idea that low-energy neutrinos change flavour as they propagate through a vacuum, and that the oscillations of high-energy neutrinos are enhanced by their interactions with electrons.

The new results also shed light on a long-standing paradox in solar physics, which arises because the chemical composition of the Sun is not well established. The most-recent complete spectroscopic determinations of the Sun's metallicity[5] (the abundance of all solar elements heavier than helium) yielded a value that is 35% lower than older spectroscopic results[6]. Intriguingly, when numerical models of the solar interior are constructed using the lower value of metallicity as a constraint, the simulated properties are at odds with our knowledge of the Sun's interior structure (which is well characterized by helioseismological studies[7] that analyse oscillations

produced by waves that propagate through the Sun's interior). But when the older (higher) metallicity values are used, the simulations reproduce solar properties very well. This is known as the solar abundance problem, and calls into question the validity of the present models of stellar evolution, or of spectroscopic methods for determining the Sun's composition, or both.

However, the relative contributions of the three different reaction sequences in the *pp*-chain, determined from the Borexino experiment, can be used to infer the temperature in the solar core — a region that is poorly mapped by helioseismological studies. The Borexino findings hint at a core temperature that is consistent with predictions from models that assume high solar metallicity. That said, the results are not yet precise enough to provide a definite answer to the solar abundance problem, because neutrino fluxes predicted by both the high- and low-metallicity solar models are compatible with the new results.

Nevertheless, the Borexino experiment might provide a definite answer in the future. About 1% of the Sun's nuclear energy is produced through chains of nuclear reactions known as CNO cycles[8]. These cycles are catalysed by the presence of carbon, nitrogen and oxygen, and so their efficiency depends linearly on solar metallicity. If the neutrino fluxes associated with CNO cycles could be measured, then the abundances of these elements in the solar core could be determined.

Such measurements have proved difficult at Borexino so far, because of background radiation produced by the radioactive decay of bismuth-210 (which forms from the decay of uranium-238, an isotope present in tiny quantities in all matter in the Solar System). Modifications to the vessel that holds the liquid scintillator have now been made[9] that should address this issue. The detection of CNO neutrinos would not only allow the Sun's metallicity to be determined, but would also provide direct evidence that CNO cycles occur in nature. This is important, because CNO cycles are thought to be the main mechanism by which stars more massive than the Sun generate energy[8].

Another major issue in astrophysics is the proposed existence of non-standard mechanisms for the production or loss of energy in stars[10]. If such a mechanism exists, there will be an imbalance between the solar production rate of nuclear energy and luminosity (the total amount of energy radiated as photons from the Sun's surface). The precision with which the power generated by nuclear reactions in the Sun can be measured would need to be increased tenfold to 1% to allow tests of such non-standard particle physics. Such precision may be out of reach for Borexino, but it might be possible in future large-scale neutrino and dark-matter detectors. ■

Aldo Serenelli *is in the Astrophysics and Planetary Sciences Department, Institute of Space Sciences (CSIC), and at the Institut d'Estudis Espacials de Catalunya, Bellaterra 08193, Spain.*
e-mail: aldos@ice.csic.es

1. The Borexino Collaboration. *Nature* **562**, 505–510 (2018).
2. Bergström, J. *et al. J. High Energ. Phys.* **3**, 132 (2016).
3. Mikheyev, S. P. & Smirnov, A. Y. *Yadernaya Fiz.* **42**, 1441–1448 (1985).
4. Wolfenstein, L. *Phys. Rev. D* **17,** 2369 (1978).
5. Asplund, M., Grevesse, N., Sauval, A. J. & Scott, P. *Annu. Rev. Astron. Astrophys.* **47**, 481–522 (2009).
6. Grevesse, N. & Sauval, A. J. *Space Sci. Rev.* **85**, 161–174 (1998).
7. Basu, S. & Antia, H. M. *Phys. Rep.* **457**, 217–283 (2008).
8. Haxton, W. C., Hamish Robertson, R. G. & Serenelli, A. M. *Annu. Rev. Astron. Astrophys.* **51,** 21–61 (2013).
9. Smirnov, O. Data available at https://doi.org/10.5281/zenodo.1286860 (2018).
10. Raffelt, G. G. *Annu. Rev. Nucl. Particle Sci.* **49**, 163–216 (1999).

DEVELOPMENTAL BIOLOGY

How to lose your inheritance

In developing embryos, molecular and physical differences divide the cells that will form eggs or sperm and those that will form the body. The mouse protein OTX2 directs this decision by blocking reproductive-cell fate.

DIANA J. LAIRD

In an ultimate act of family planning, the cells destined to contribute to the next generation of an organism are set aside early in embryonic development. It is unclear why primordial germ cells (PGCs), the precursors of eggs and sperm, are established so early in the development of many multicellular organisms. This process of establishing the germ line involves both preventing a non-reproductive-cell (somatic-cell) fate and activating a cellular state known as pluripotency — the ability to give rise to the many different cell types in the body. Understanding how the germ line forms is a key requirement for the goal of generating healthy eggs and sperm *in vitro* for fertility treatment in the clinic. Work in this area has focused mainly on identifying proteins that specify germline fate, but comparatively little is known about why somatic cells do not acquire such a fate. Writing in *Nature*, Zhang *et al.*[1] reveal that the formation of PGCs in mice can be blocked by a protein called OTX2.

In many animals, specification of the germ line operates like a hereditary monarchy in which, like the direct transmission of the crown jewels from one generation to the next, the passage of molecular components in the cytoplasm down the generations determines the cells that will form PGCs. However, some animals, including salamanders, crickets, mice and possibly humans, take a different approach. In the early mouse embryo, designation of the germ line occurs as a result of cells being in the right place at the right time, rather than inheriting the species' crown jewels. In this inductive fate-determination scenario, cells of a cylindrically shaped region of pluripotent cells known as the epiblast are coaxed into adopting a PGC fate by signals from supporting extraembryonic tissues adjacent to the embryo.

In this mode of germline formation, the instructive cues that travel between cells include proteins of the Wnt and BMP families[2]. PGCs normally form at a predictable location in the epiblast (Fig. 1). However, *in vivo* grafting experiments in mouse embryos revealed that cells from elsewhere in the epiblast have the capacity to become PGCs if they are transplanted to that location[3]. The search for the components that drive germline fate in epiblast cells in response to the 'kingmaker' BMP proteins identified the transcription-factor proteins BLIMP1, PRDM14 and AP2γ (refs 4, 5). This trio of proteins not only drives the expression of genes required to make the germ line, but also blocks the expression of genes associated with a somatic-cell fate[4,5].

The process of PGC development can be recapitulated *in vitro,* starting from mouse embryonic stem (ES) cells that can be induced to form epiblast-like cells. If these epiblast-like cells are exposed to BMPs and certain other factors, then as many as 13.5% of the cells form PGCs[6]. However, if such epiblast-like cells are engineered to express BLIMP1, PRDM14 and AP2γ, more than 30% form PGCs without the requirement[4] for BMPs. Yet knowing that a particular pathway can drive the formation of PGCs doesn't answer the questions of whether the default pathway of cellular differentiation in embryonic development is to form germline or somatic cells, or whether all of the cells in the epiblast are equally capable of becoming germline cells. Both matters have implications for our understanding of the evolution of multicellularity, as well as for our ability to generate healthy eggs or sperm from stem cells for clinical applications.

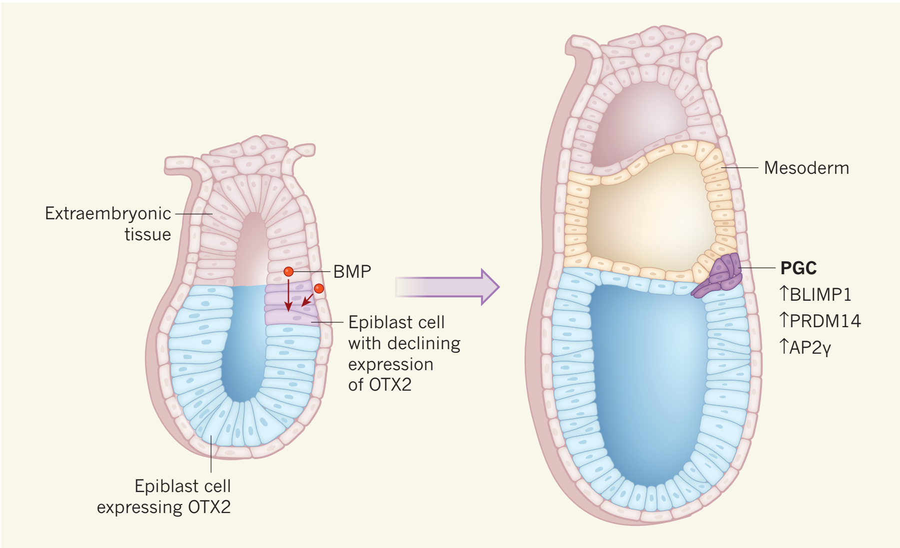


Figure 1 | Expression of the transcription factor OTX2 declines in cells that will form primordial germ cells. Zhang *et al.*[1] report how primordial germ cells (PGCs), the precursors of egg and sperm cells, form *in vivo* in mice from cells in an embryonic region called the epiblast. The formation of PGCs requires a signal from BMP proteins in the adjacent extraembryonic region[2]. The authors report that there is a fall in OTX2 expression in the epiblast region where PGCs will subsequently arise. The drop in OTX2 expression in these cells is followed by a rise in the expression of other transcription factors — BLIMP1, PRDM14 and AP2γ — that are involved the formation of PGCs[4,5], which form in an area adjacent to a layer of embryonic cells called the mesoderm. The authors' results indicate that OTX2 blocks the establishment of PGCs.

Zhang and colleagues studied OTX2, a transcription factor known to be involved in the development of the nervous system. *In vitro* studies had revealed[7] that OTX2 promotes the transition of mouse ES cells into epiblast-like cells. Zhang *et al.* scrutinized the expression of OTX2 in mouse stem cells that were differentiating into germline cells *in vitro*. They found that, in epiblast-like cells progressing towards PGC formation, OTX2 levels decline 12–24 hours before BLIMP1, PRDM14 and AP2γ are upregulated. The authors' *in vivo* analysis in mice produced similar results (Fig. 1). Considered together, these patterns of expression make sense: the level of OTX2 is low in PGCs and ES cells, both of which exist in a state of pluripotency termed naive pluripotency, which is defined as the ability to give rise to all types of cell in the body. By contrast, the level of OTX2 is high in the epiblast and in ES-cell-derived epiblast-like cells, which are considered to be more restricted in the number of cell lineages they can form.

The timing of the OTX2 decline before the transition from epiblast-cell to PGC fate raises the question of whether OTX2 might prevent cells from transitioning into PGCs. To investigate this, Zhang *et al.* studied the formation of PGCs in the absence of OTX2. Compared with the situation for wild-type cells, elevated numbers of PGCs were found if embryos lacked the *Otx2* gene, or if epiblast-like cells grown *in vitro* were *Otx2*-deficient. Conversely, if epiblast-like cells were engineered to express higher than usual levels of OTX2, the formation of PGCs was prevented.

Moreover, the authors found that the acquisition of PGC fate in the absence of *Otx2* could be accomplished *in vitro* without the usual requirements of BLIMP1 and signalling molecules called cytokines. This was unexpected because *in vivo* and *in vitro* experiments[8,9] have indicated that PGC production requires BLIMP1. How is the requirement for BLIMP1 in germline-cell formation bypassed? Given that OTX2 promotes a form of pluripotency in the epiblast that is more restricted than that of naive pluripotency[7], perhaps the primary function of BLIMP1 is to create a naive state of pluripotency for newly forming PGCs.

The authors' results are consistent with a model in which BMPs, possibly acting through Wnts, cause the level of OTX2 to decline. This reduction in OTX2 is one of the earliest known steps that determine whether an epiblast cell will acquire a germ-cell or somatic-cell fate. In an intriguing parallel to the system in fruit flies and worms, in which transcription is repressed altogether and cytoplasmic 'crown jewels' are inherited[10], the first commitment to a PGC fate in mice is now revealed to be a process in which cells are defined by what they are not. Extinguishing the transcription factor OTX2 prevents it from driving gene expression associated with the more-restricted state of pluripotency of the epiblast[7], instead making way for the naive pluripotency needed for PGCs.

The authors show that the absence of OTX2 drives PGC formation, but it remains to be determined whether these PGCs differentiate normally and form fully functional eggs and sperm. Might the levels of OTX2 affect the fitness of germline cells? For example, would germline cells that have low levels of OTX2 be less likely to successfully contribute to reproduction than cells in which OTX2 has been completely extinguished? Conversely, does the level of OTX2 affect the fitness of somatic cells?

Many other questions remain to be answered. How do BMPs and Wnts cause a decline in OTX2 levels? What is the regulatory relationship between OTX2 and other transcription factors that act in developing PGCs? In the development of the mouse eye[11], OTX2 drives the expression of BLIMP1 in a process linked to a switch between two types of cell fate, yet in the PGCs, BLIMP1 seems to function after OTX2 levels have declined.

In the developing ear of the fruit fly *Drosophila melanogaster*[12], the transcription of OTX2 is directly regulated by the transcription factor N-myc. The known roles of Myc-family genes in metabolism and in a process known as cell competition[13] suggest that even minor variations in OTX2 levels in the epiblast might reflect a biologically meaningful variation in the fitness of cells that will go on to form the somatic-cell lineages or the germ line. In other words, if it turns out that the levels of Myc correlate with the health of an epiblast cell and its ability to outcompete its neighbouring cells, and if Myc controls the level of OTX2 in mice, then OTX2 might have a role in the fitness of epiblasts, and perhaps in that of PGCs.

OTX2 could provide an avenue for investigating the cell-fate decision between forming germline and somatic cells, and might offer ways of improving PGC differentiation in *in vitro* approaches. If this mechanism of OTX2-mediated regulation of PGC fate in mice is evolutionarily conserved, then perhaps similar progress might be made in such studies of human cells. ■

Diana J. Laird *is in the Department of Obstetrics, Gynecology and Reproductive Science, University of California, San Francisco, and at the Eli and Edythe Broad Center for Regeneration Medicine and Stem Cell Research, San Francisco, California 94143, USA.*
e-mail: diana.laird@ucsf.edu

1. Zhang, J. *et al. Nature* **562**, 595–599 (2018).
2. Ohinata, Y. *et al. Cell* **137**, 571–584 (2009).
3. Tam, P. P. L. & Zhou, S. X. *Dev. Biol.* **178**, 124–132 (1996).
4. Nakaki, F. *et al. Nature* **501**, 222–226 (2013).
5. Magnúsdóttir, E. *et al. Nature Cell Biol.* **15**, 905–915 (2013).
6. Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. *Cell* **146**, 519–532 (2011).
7. Acampora, D., Di Giovannantonio, L. G. & Simeone, A. *Development* **140**, 43–55 (2013).
8. Ohinata, Y. *et al. Nature* **436**, 207–213 (2005).
9. Vincent, S. D. *et al. Development* **132**, 1315–1325 (2005).
10. Strome, S. & Lehmann, R. *Science* **316**, 393–393 (2007).
11. Mills, T. S. *et al. PLoS ONE* **12**, e0176905 (2017).
12. Vendrell, V. *et al. Development* **142**, 2792–2800 (2015).
13. Clavería, C. *et al. Nature* **500**, 39–44 (2013).

This article was published online on 3 October 2018.

CONSERVATION

Parrot patrol

Understanding the movements of animals can help to focus conservation efforts on key regions of habitat. A standard approach to tracking small animals is to tag them with a tiny radio transmitter and monitor the emitted signals using a hand-held device. But this can prove challenging when following highly mobile animals across difficult terrain.

Writing in *Science Robotics*, Cliff *et al.* report their analysis of the use of autonomous aerial vehicles to track tagged birds (O. M. Cliff *et al. Sci. Robot.* **3**, eaat8409; 2018). They monitored wild swift parrots (*Lathamus discolor*, pictured), an endangered Australian species. The authors report that these drones can estimate birds' positions as rapidly as can humans experienced at using the standard manual-tracking method. **Mary Abraham**

SIMON BENNETT/BIA/MINDEN/GETTY

OPTICAL PHYSICS

Precise control of infrared polarization

A natural material has been discovered that exhibits an extreme optical property known as in-plane hyperbolicity. The finding could lead to infrared optical components that are much smaller than those now available. SEE LETTER P.557

THOMAS G. FOLLAND & JOSHUA D. CALDWELL

Hyperbolic materials are highly reflective to light along a certain axis and reflective to light along a perpendicular axis. Typically, one of these axes is in the plane of the material and the other is out of the plane. A material in which both of these axes are in the plane would enable, for example, the manufacture of ultrathin waveplates — optical components that alter the polarization of incident light. Moreover, the reflective behaviour of this material would allow light to be confined and manipulated at extremely small scales (less than one-hundredth the wavelength of the light). On page 557, Ma *et al.*[1] report the existence of such in-plane hyperbolicity in the natural material molybdenum trioxide.

Many crystals exhibit birefringence, in which their refractive index — a measure of the speed of light in a material — is different along different axes. This property can be used to manipulate the polarization of incident light. The crystal size that is required to achieve sufficient polarization control for practical applications is directly proportional to the wavelength of the incident light and to the strength of the birefringence. Consequently, for light in the mid- to far-infrared regions of the electromagnetic spectrum (with wavelengths of 3–300 micrometres), the crystals typically need to be a few millimetres thick[2]. To overcome this requirement, a potential solution is to consider materials that exhibit hyperbolicity, which is an extreme form of birefringence.

Hyperbolicity was originally thought to exist only in artificial materials consisting of integrated reflective and transparent domains[3]. But in 2014, it was observed in the natural material hexagonal boron nitride[4,5]. The reflective behaviour of both this material and molybdenum trioxide is derived from crystal-lattice vibrations, known as optical phonons, that oscillate in a highly anisotropic (direction-dependent) way. These phonons have relatively long lifetimes (in excess of a picosecond; 1 ps is 10^{-12} s), which strongly suppresses the absorption of light by the material[6]. Since the discovery of hyperbolicity in hexagonal boron nitride, a broad array of natural hyperbolic materials has been identified[7].

Preliminary investigations of molybdenum trioxide were reported earlier this year[8] and showed the existence of hyperbolicity for long-wave infrared light (with wavelengths of 8–14 μm). Ma and colleagues have now demonstrated and characterized in-plane hyperbolicity for the same spectral range. They used this property to confine light to dimensions substantially smaller than its wavelength, through the formation of hybrid light–matter excitations called hyperbolic phonon polaritons. The authors report lifetimes for such polaritons of up to 20 ps, which is about ten times longer than the best values reported for hexagonal boron nitride[9].

Because the crystal structure of molybdenum trioxide is highly anisotropic, all three crystal axes, which define the edges of the crystal's unit cell, have different lengths. Consequently, there is a large difference in the phonon energies associated with these axes and therefore in the corresponding refractive indices — resulting in a birefringence of about 0.31. It should be noted that, earlier this year, a similarly large in-plane birefringence of 0.76 was reported in the natural material barium titanium sulfide for mid-infrared to long-wave infrared light[10]. However, hyperbolicity was not observed for this material.

The in-plane hyperbolicity of molybdenum trioxide offers opportunities to replace conventional optical components with ones that are much smaller. In particular, using the large in-plane birefringence of this material (or of barium titanium sulfide), infrared waveplates could be constructed from thin slabs that have

NATURE

50 Years Ago

Clearance by Carp
Britain spends about £2.5 million a year on removing water weeds from inland waterways, and in 1964 the British River Authorities spent on average £69 per mile on weed clearance. In an attempt to cut these costs the Ministry of Agriculture, since 1964, has been experimenting with the grass carp, a fish native to China, but widely cultivated … as a means of biological control of water weed. Although the experiments are still at a very early stage, no major snags have occurred so far and the ministry's Salmon and Fresh Water Fisheries Laboratory is still optimistic that the method may work.
From *Nature* 26 October 1968

100 Years Ago

There is a general belief that it is a relatively easy problem to estimate a person's intelligence by looking at him; and teachers, physicians, and employers are often compelled to make judgments as to the intelligence of a given person with no more data than can be obtained from a rapid survey of his appearance … In the *Psychological Review* … Mr. R. Pinter gives the results of an investigation he made for the purpose of testing the trustworthiness of these judgments. The author chose twelve photographs of children varying in intelligence from proved feeble-mindedness to unusually great ability and asked groups of people to arrange the photographs in order of merit for intelligence. His groups consisted of physicians, psychologists, teachers, and miscellaneous people. He found that … on the judgment of no one group or of no one person could any reliance be placed … The author concludes that … these haphazard judgments are too untrustworthy to be of practical value.
From Nature 24 October 1918

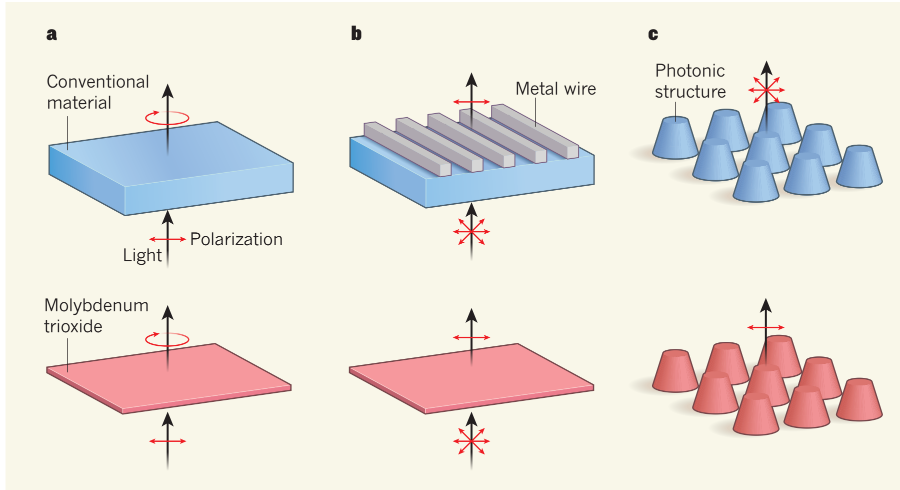


Figure 1 | Manipulating infrared polarization. Ma *et al.*[1] show that the material molybdenum trioxide can be used to precisely control the polarization of infrared light. **a**, Optical components known as waveplates can convert linearly polarized light into circularly polarized light. In the infrared, a waveplate made of a conventional material requires a thickness in excess of 1 millimetre. This material could be replaced with a thin slab of molybdenum trioxide, with a thickness on the order of tens of micrometres. **b**, Components called polarizers can convert unpolarized light (in which the polarization points in all directions) into linearly polarized light. In the infrared, polarizers made from conventional materials typically need to be thick and use a large grid of metal wires. Such a structure could be replaced with a thin film of molybdenum trioxide that requires essentially no fabrication. **c**, Nanoscale photonic structures made from conventional materials can emit unpolarized infrared light. But if molybdenum trioxide were used, linearly polarized emission could be achieved.

thicknesses on the order of tens of micrometres (Fig. 1a). Such waveplates could operate in the long-wave infrared, for which commercial waveplates are not widely available and have thicknesses in excess of 1 mm.

Furthermore, using the material's in-plane hyperbolicity, polarizers — components that extinguish undesired polarizations of incident light — could be made from simple 1-μm-thick films (Fig. 1b). Previously, polarizers needed to be thicker and typically required a large grid of metal wires to be patterned on their surface. The remarkable properties of molybdenum trioxide could therefore greatly reduce both the size and the cost of optical components, offering broad applicability in thin, compact infrared devices.

Beyond conventional optics, the properties of molybdenum trioxide could lead to advances in the realm of nanophotonics, which focuses on confining light to nanoscale dimensions. In the long-wave infrared, where the hyperbolicity of this material is observed, nanoscale light confinement necessarily implies defeating the diffraction limit — the usual restriction that light cannot be squeezed into dimensions much smaller than its wavelength. Molybdenum trioxide can beat this limit and, as a result, presents opportunities for producing improved infrared-emitting devices.

For instance, heating nanoscale photonic structures made from materials that can support polaritons can produce light of one or more specific frequencies — rather than light of a broad range of frequencies that that emitted by, for example, conventional light bulbs. Such structures provide an optical source that is akin to light-emitting diodes, but that can be designed to operate anywhere in the infrared. The emitted light from these photonic structures is usually unpolarized (Fig. 1c). It is only through the use of materials that exhibit in-plane hyperbolicity that light of a single, pure polarization can be generated.

Finally, hyperbolic materials such as molybdenum trioxide could serve as the basis for hyperlenses — lenses that produce magnified images of objects smaller than the wavelength of the imaging light. They could also be used in heterostructures (structures in which layers of different materials are combined) to make nanophotonic components that have controllable properties[11,12].

Ma and colleagues have demonstrated that, once again, nature has more in store for us than we thought. The future of nanophotonics was once considered to be in the realization of artificial materials, but this study and others in the past few years have demonstrated that, in many cases, the best approach for finding advanced materials is to look among the vast array of natural materials. The results of these studies offer substantial advances in the fields of infrared optics and nanophotonics that could enable infrared imaging and detection to become as ubiquitous as its visible counterpart — a vision that would enable imaging through smoke for first responders, near-instant medical diagnostics and enhanced chemical spectroscopy. ■

Thomas G. Folland *and* **Joshua D. Caldwell** *are in the Department of Mechanical Engineering, Vanderbilt University, Nashville, Tennessee 37212, USA.*
e-mails: thomas.g.folland@vanderbilt.edu;

josh.caldwell@vanderbilt.edu

1. Ma, W. *et al. Nature* **562**, 557–562 (2018).
2. Suslikov, L. M., Gadmashi, Z. P., Kovach, D. Sh. & Slivka, V. Yu. *Opt. Spectrosc.* **53**, 283–287 (1982).
3. Poddubny, A., Iorsh, I., Belov, P. & Kivshar, Y. *Nature Photon.* **7**, 948–957 (2013).
4. Dai, S. *et al. Science* **343**, 1125–1129 (2014).
5. Caldwell, J. D. *et al. Nature Commun.* **5**, 5221 (2014).
6. Caldwell, J. D. *et al. Nanophotonics* **4**, 44–68 (2015).
7. Korzeb, K., Gajc, M. & Pawlak, D. A. *Optics Express* **23**, 25406–25424 (2015).
8. Zheng, Z. *et al. Adv. Mater.* **30**, 1705318 (2018).
9. Giles, A. J. *et al. Nature Mater.* **17**, 134–139 (2018).
10. Niu, S. *et al. Nature Photon.* **12**, 392–396 (2018).
11. Li, P. *et al. Nature Mater.* **15**, 870–875 (2016).
12. Folland, T. G. *et al. Nature Commun.* https://doi.org/10.1038/s41467-018-06858-y (2018).

SUSTAINABILITY

Transforming the global food system

Can the predicted rise in global food demand by 2050 be met sustainably? A modelling study suggests that a combination of interventions will be needed to tackle the associated environmental challenges. SEE ARTICLE P. 519

GÜNTHER FISCHER

The global population in 2010 was estimated to be 6.9 billion people, and by 2050 is predicted to reach between 8.5 billion and 10 billion people[1]. This increase would bring a corresponding rise in food demand that would affect the environmental toll that food production exerts on the planet. On page 519, Springmann *et al.*[2] report their analysis of the environmental pressures that would arise in a projected scenario for the global food system in 2050. They also modelled the effects of implementing approaches to lessen the environmental consequences of food production.

Food security has long been a challenge for human societies, and is a pressing global issue. Indeed, many targets related to this area are part of the United Nations' Sustainable Development Goals[3], which include eradicating hunger, ending poverty and combating climate change. Achieving a sustainable global food system clearly requires progress on social, economic and environmental fronts.

Springmann and colleagues built a model to assess the projected global demand for agricultural products by 2050 on a country-by-country basis, given the expected changes in population, income levels and dietary preferences by that time. It has been predicted[4] that global income in 2050 will be 3–4 times higher than it was in 2010. The authors' projections of future food consumption were based on established statistical associations between food demands and changes in income or population. These predict that, by 2050, there will be less undernutrition, a shift towards greater global consumption of livestock-based products and a fairly constant intake of staple crops per person.

The authors assessed predicted global environmental impacts for the projected food production by mid-century. They focused on five environmental pressures: the greenhouse-gas emissions associated with agricultural production; the use of land for crop production, given the associated consequences (such as carbon or biodiversity losses) that might accompany land-use changes; the demand for water to irrigate crops; and the application of either nitrogen- or phosphorus-based fertilizers, respectively. It is important to consider fertilizers because of the greenhouse-gas emissions that are linked to their use, and the possibility that they might contaminate soils or aquatic ecosystems.

Springmann *et al.* compared the projected environmental impacts in 2050 to a proposed set of planetary boundaries thought to represent safe operating limits for human activities[5]. For example, the boundary set by the authors for agricultural greenhouse-gas emissions was established in relation to the threshold necessary to keep global warming at a level of 2 °C above pre-industrial levels. However, their limit for emission levels is less stringent than the limit needed to achieve the 1.5 °C target set in the United Nations Framework Convention on Climate Change Paris Agreement of 2015, which was analysed in a recent report[6] by the Intergovernmental Panel on Climate Change. This report details how limiting warming to 1.5 °C rather than to 2 °C above pre-industrial levels would reduce the climate-related risks to health, livelihoods, food security and water supply. On the basis of current food yields and agricultural practices, Springmann and colleagues estimate that, between 2010 and 2050, the environmental impacts of the food system could increase by between 50% and 92% and reach levels that exceed the proposed boundaries[5] for planetary stability.

The researchers went on to assess the effect of possible interventions that could reduce these environmental pressures. These measures relate to managing food demand and raising food-production efficiency in terms of three broad intervention categories.

One intervention category concerns improvements in agricultural technologies and resource management. These could enhance production efficiency and increase crop yields per unit of land, given a particular water and nutrient input. Another category was dietary changes, whereby individuals might limit their meat consumption and move towards plant-based foods. Meat production usually requires a more intensive and environmentally damaging mode of production than that needed for plant-based food[7]. Moreover, limiting meat and sugar consumption and eating fruit and vegetables is aligned with nutrition guidelines for a healthy diet[8]. The third category the authors considered was

BEN NELMES/REUTERS

Figure 1 | Discarded food waste in British Columbia, Canada. This food did not reach consumers.

reduction of food-chain waste from field to plate. It is estimated that up to one-third of food doesn't reach the market (Fig. 1) or is discarded after purchase[9]. Reducing this waste would increase food availability without the need for extra food production.

Springmann and colleagues conclude that an intervention in only one of the three categories they analysed would not achieve planetary sustainability across all five of the environmental domains that they assessed. Instead, a bundle of interventions in all three categories would be needed to ensure that the global food system could be sustainably supported by the planet in 2050. They found that the projected greenhouse-gas emissions from agriculture would not be supportable unless global meat consumption was reduced. They also report that the expansion of cropland and water use would be best counteracted by improvements in agricultural technologies and management approaches that bring farming yields closer to the maximum yield efficiency that is ecologically possible. In addition, their analysis indicates that achieving fertilizer-use reduction would require a combination of measures that improve farming practices and decrease food demand.

There are some caveats regarding Springmann and colleagues' scenarios. For example, they did not take climate-change effects into account in their projections of future agricultural production, and such impacts should be a priority for future analysis. Also, the authors' analysis did not consider the world's grassland areas, even though they represent more than double the area of global cropland[10]. These grassland areas should be considered when setting planetary boundaries for land use. Moreover, Springmann and colleagues' study analyses only the environmental impacts of cropland-based food production — it doesn't assess how to balance these impacts with those in sectors such as energy, transport or industry.

Nevertheless, the authors' analysis is valuable and informative for the discussion about how to achieve a sustainable food system that meets future needs, even if some of the planetary-boundary values they used have large uncertainty ranges[11]. In addition, any proposed interventions should not be implemented using a one-size-fits-all approach. Instead, any regulatory frameworks and incentives will need to be tailored to the needs of a given region, whether this means investments in education, health-service access, land-use regulations or water allocation, for example.

Springmann and colleagues also did not address certain key issues that are needed to develop a resilient agricultural system. The rights of access to land and natural resources, and the long-term security of those rights, is needed to motivate investments by farmers. Farmers could also be helped by improvements in transport, finance and communication infrastructure that enable them to access advanced technologies, minimize their production risks and target their production for local or international markets.

A recent report[12] by the Food and Agriculture Organization of the United Nations concludes that environmental sustainability and food security can go hand in hand by 2050, but that substantial investments are needed to transform the global food system. Political and public commitment will be essential to ensure increases in budgets for the development of international agriculture.

Food demand and food production are two sides of the global food-system equation. Springmann and colleagues' work provides a timely warning that interventions will be needed in both domains to achieve food security in the future, and to ensure that the environmental impacts of the food-production system remain within boundaries that Earth can sustain. ■

Günther Fischer *is at the International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria.*
e-mail: fischer@iiasa.ac.at

1. Samir, K. C. & Lutz, W. *Glob. Environ. Change* **42**, 181–192 (2017).
2. Springmann, M. *et al. Nature* **562**, 519–525 (2018).
3. United Nations. *Transforming our World: The 2030 Agenda for Sustainable Development* (UN, 2015).
4. Dellink, R., Chateau, J., Lanzi, E. & Magné, B. *Glob. Environ. Change* **42**, 200–214 (2017).
5. Rockström, J. *et al. Nature* **461**, 472–475 (2009).
6. Intergovernmental Panel on Climate Change. *Global Warming of 1.5 °C* (IPCC, 2018).
7. Tilman, D. & Clark, M. *Nature* **515**, 518–522 (2014).
8. World Health Organization. *Healthy Diet* Fact Sheet No. 394 (WHO, 2018).
9. Food and Agriculture Organization of the United Nations. *Global Food Losses and Food Waste: Extent, Causes and Prevention* (FAO, 2011).
10. Food and Agricultural Organization of the United Nations. FAOSTAT 2018. available at http://www.fao.org/faostat/en/#data/RL
11. Jaramillo, F. & Destouni, G. *Science* **348**, 1217 (2015).
12. Food and Agriculture Organization of the United Nations. *The Future of Food and Agriculture: Alternative Pathways to 2050* (FAO, 2018).

MATERIALS SCIENCE

The war on fake graphene

The material graphene has a vast number of potential applications — but a survey of commercially available graphene samples reveals that research could be undermined by the poor quality of the available material.

PETER BØGGILD

Graphite is composed of layers of carbon atoms just a single atom in thickness, known as graphene sheets, to which it owes many of its remarkable properties. When the thickness of graphite flakes is reduced to just a few graphene layers, some of the material's technologically most important characteristics are greatly enhanced — such as the total surface area per gram, and the mechanical flexibility of the individual flakes. In other words, graphene is more than just thin graphite. Unfortunately, it seems that many graphene producers either do not know or do not care about this. Writing in *Advanced Materials*, Kauling *et al.*[1] report a systematic study of graphene from 60 producers, and find that many highly priced graphene products consist mostly of graphite powder.

Imagine a world in which antibiotics could be sold by anybody, and were not subject to quality standards and regulations. Many people would be afraid to use them because of the potential side effects, or because they had no faith that they would work, with potentially fatal consequences. For emerging nanomaterials such as graphene, a lack of standards is creating a situation that, although not deadly, is similarly unacceptable.

One of the most well-established methods for producing graphene for commercial applications is liquid-phase exfoliation[2] (LPE) — a process that involves milling graphite into a powder, and separating the particles into tiny flakes by applying mechanical forces in a liquid. Those precious flakes that contain just a few layers of graphene are then separated from the rest (Fig. 1). Graphene produced in this way has a huge number of potential applications, including battery technology, composite materials and solar cells. The LPE of graphite was first achieved using sonication to produce the flakes[3], and later work showed that even a kitchen blender[4] can be used to create violent turbulent forces that pull graphene sheets apart without destroying them.

But how thin must graphite flakes be to behave as graphene? A common idea, backed up[5] by the International Organization for Standardization (ISO), is that flakes containing more than ten graphene layers are basically graphite. This seemingly arbitrary threshold has some basis in physics, as Kauling *et al.* note. For example, thermodynamic considerations dictate that each layer of atoms in a flake of ten or fewer layers behaves as an individual graphene crystal at room temperature.

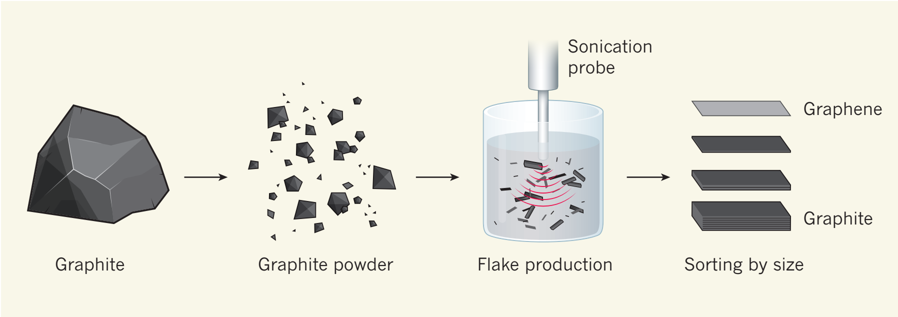


Figure 1 | Liquid-phase exfoliation of graphene. Most commercially available bulk graphene is made by milling graphite into powder, and then subjecting the resulting particles to mechanical forces in a liquid solution to separate the powder into flakes, for example, by using sonication; flakes not shown to scale. The flakes are then sorted according to their size and thickness. Kauling *et al.*[1] analysed commercially available graphene from 60 providers, and found that the majority of the samples contained less than 10% of graphene (flakes that contain fewer than ten layers of carbon atoms[5]). The rest is essentially just graphite powder. (Adapted from ref. 1.)

Moreover, the rigidity of flakes scales with the cube of layer thickness, which means that thin graphene flakes are orders of magnitude more flexible than thicker graphite flakes.

So size really matters: depending on the practical application, graphene and graphite powders can give entirely different results. Without clear standards by which to determine the quality of commercially available graphene, companies and researchers risk wasting time and money doing research on graphite powder disguised as expensive, high-grade graphene. This would stunt the development of graphene technology, harming serious graphene producers and application developers alike.

But are these concerns truly warranted? In a study aimed at answering this question, Kauling *et al*. established a systematic test protocol based on an arsenal of well-established methods for characterizing graphene, and then used the protocol to benchmark 60 graphene products from different producers, a daunting task. The results showed that the statistical distributions of the key material indicators — such as the size, structural integrity and purity of the graphene — varied greatly. Shockingly, the study revealed that less than 10% of the material in most of the products consisted of graphene composed of ten or fewer layers. None of the products tested contained more than 50% of such graphene, and many were heavily contaminated, most likely with chemicals used in the production process.

It seems that the high-profile scientific discoveries, technical breakthroughs and heavy investment in graphene have created a Wild West for business opportunists: the study shows that some producers are labelling black powders that mostly contain cheap graphite as graphene, and selling them for top dollar. The problem is exacerbated because the entry barrier to becoming a graphene provider is exceptionally low — anyone can buy bulk graphite, grind it to powder and make a website to sell it on.

Unless common standards and test protocols are introduced, there is a great risk of dropping the ball at the worst possible time. Dozens of emerging applications for graphene are closely linked to some of society's grand challenges: health, climate, renewable energy and sustainability. Some of these applications might never leave the starting block if the early development is based on 'fake graphene'.

Kauling and colleagues' article is therefore a much-needed wake-up call for graphene producers, buyers and researchers to agree on and to adhere to sound standards: a transparent graphene market would benefit everyone, except perhaps unscrupulous vendors. The first steps towards this have already been taken with the ISO's graphene vocabulary[5] (a document that defines standard terminology for describing graphene) and the UK National Physical Laboratory's helpful Good Practice Guide for graphene characterization[6]. Now it's time to push on.

It should be noted that Kauling and co-workers' study does not cover all the types of bulk graphene on the market[7]. Moreover, although the authors analysed an impressive number of LPE-manufactured products, they could have eliminated any accusations of potential bias by specifying the criteria they used to select the products for analysis. It is also possible that they unintentionally missed high-quality graphene sold by a few excellent producers. And, as the researchers mention, different applications generally make use of different characteristics of graphene — which makes it difficult to come up with a universal metric of quality.

Nevertheless, the work is a timely and ambitious example of the rigorous mindset needed to make rapid progress, not just in graphene research, but in work on any nanomaterial entering the market. To put it bluntly, there can be no quality without quality control. ■

Peter Bøggild *is in the Department of Micro- and Nanotechnology and in the Center for Nanostructured Graphene, Technical University of Denmark, 2800 Kongens Lyngby, Denmark.*
e-mail: peter.boggild@nanotech.dtu.dk

1. Kauling, A. P. *et al. Adv. Mater.* **2018**, 1803784 (2018).
2. Bonaccorso, F. *et al. Mater. Today* **15**, 564–589 (2012).
3. Hernandez, Y. *et al. Nature Nanotechnol.* **3**, 563–568 (2008).
4. Paton, K. R. *et al. Nature Mater.* **13**, 624–630 (2014).
5. ISO. *Nanotechnologies-Vocabulary-Part 13: Graphene and Related Two-Dimensional (2D) Materials* www.iso.org/standard/64741.html (2017).
6. National Physical Laboratory. *Characterisation of the Structure of Graphene* (NPL, 2017).
7. Xia, Z. Y. *et al. Adv. Funct. Mater.* **23**, 4684–4693 (2013).

This article was published online on 8 October 2018.

NEUROSCIENCE

Senescence mediates neurodegeneration

Aggregation of the protein tau is implicated in neurodegenerative diseases in humans. It emerges that eliminating a type of damaged cell that no longer divides can prevent tau-mediated neurodegeneration in mice. SEE LETTER P.578

JAY PENNEY & LI-HUEI TSAI

There is strong interest in understanding how neurodegeneration is affected by a cellular state called senescence, in which cells stop dividing, suppress intrinsic cell-death pathways and release pro-inflammatory molecules that can harm healthy neighbours[1,2]. On page 578, Bussian *et al.*[3] examine the role of senescent cells in a mouse model of a type of neurodegeneration that involves aggregation of the protein tau. They find that neuronal expression of mutant tau triggers senescence in glia, the support cells of the brain. Preventing the build-up of senescent glia can block the cognitive decline and neurodegeneration normally experienced by these mice.

Senescent cells are characterized by various molecular and gene-expression changes, including elevated levels of the cell-cycle inhibitor protein p16[INK4A]. Senescence can be identified by a test that stains cells blue if they

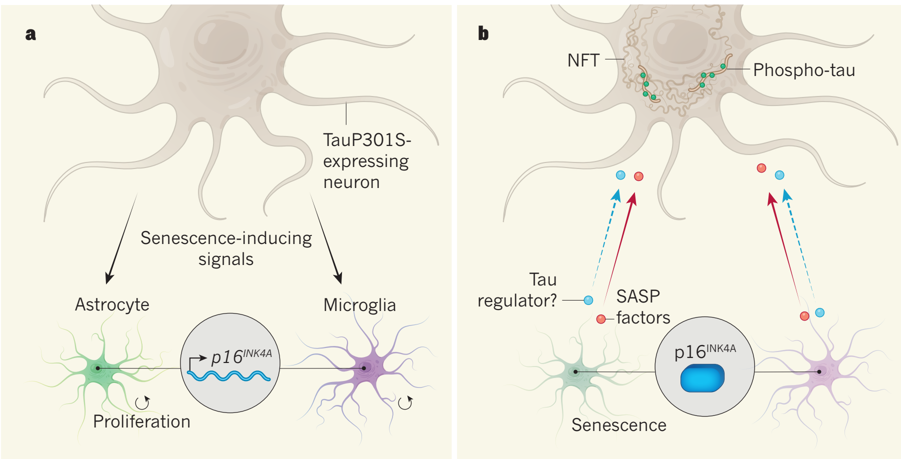


Figure 1 | Cell crosstalk in neurodegeneration. Mice that have been genetically engineered so that their neurons produce a mutant form of the protein tau (tauP301S) model some neurodegenerative diseases of humans. **a**, Bussian *et al.*[3] provide evidence that tauP301S expression leads neurons in these mice to release unknown signals that induce a cellular state called senescence in neighbouring cells. As a result, genes such as *p16[INK4A]* that are associated with senescence are activated in cells called astrocytes and microglia, which can proliferate. **b**, When senescent, astrocytes and microglia have elevated levels of p16[INK4A] protein and stop proliferating. They release a group of molecules known as senescence-associated secretory phenotype (SASP) factors that, possibly in combination with other regulators of tau, signal back to neurons. This leads to the phosphorylation of tau and its aggregation into structures called neurofibrillary tangles (NFTs) — two hallmarks of neurodegeneration.

harbour senescence-associated β-galactosidase (SA-β-Gal) — a form of the β-Gal enzyme that is active at pH 6 (in healthy cells, β-Gal is inactive at this pH)[1,4]. The cells also secrete inflammatory signalling molecules, growth factors and protease enzymes that can impair the function, and ultimately the survival, of non-senescent cells in their vicinity[1,4]. This trait is known as the senescence-associated secretory phenotype (SASP).

The gradual build-up of senescent cells contributes to ageing in multicellular organisms[1,2]. Furthermore, senescence can be induced by various cellular insults. Senescent neurons or glia have been described in people with brain injury or neurodegenerative disorders such as Parkinson's and Alzheimer's diseases[1,2,5,6]. Strategies have been developed that selectively target and eliminate senescent cells, counteracting many of the effects of ageing and age-related disorders in animal models[1,7,8]. But despite intense study, the exact effect of senescent cells in different contexts — including in neurodegeneration — remains unclear.

Bussian *et al.* set out to examine the role of senescence in neurodegeneration. They focused on the aggregation-prone neuronal protein tau, which is associated with multiple forms of neurodegeneration. For instance, a mutation in tau that changes amino-acid residue 301 from proline to serine (dubbed tauP301S) causes frontotemporal dementia[9]. And, when phosphorylated at abnormally high levels, tau forms structures called neurofibrillary tangles (NFTs) that are a hallmark of Alzheimer's disease[9].

The authors made use of mice that have been engineered to express human tauP301S in neurons, and so model human tau-mediated neurodegenerative diseases. They found elevated levels of various senescence-associated genes, including *p16[INK4A]*, in the brains of tauP301S-expressing mice compared with control animals. Using electron microscopy, the researchers examined which types of brain cell stained for SA-β-Gal in tauP301S mice. They observed no staining in neurons, but SA-β-Gal was detected in the two main types of glia — astrocytes and microglia. The group complemented their electron microscopy with an examination of senescence-associated gene expression in isolated brain-cell types. This, too, provided evidence of senescence in astrocytes and microglia, but not in neurons.

Importantly, Bussian and colleagues found that senescence-associated gene expression in tauP301S mice increased with age, but preceded NFT deposition and neurodegeneration. This suggests that the emergence of senescent cells could affect the latter two traits. To examine this possibility, the researchers eliminated senescent cells in the animals as they arose, by using a genetic tool that causes expression of a cell-death-promoting enzyme specifically in cells that produce p16[INK4A]. Removal of senescent cells prevented brain shrinkage and thinning of a cognition-related brain region called the dentate gyrus — two characteristics of tau-mediated neurodegeneration typically seen in tauP301S animals. Furthermore, cognitive function was maintained in tauP301S mice lacking senescent cells, whereas tauP301S animals in which senescent cells were retained exhibited short-term memory defects.

Perhaps more surprisingly, given that it indicates complex crosstalk between neurons and senescent glia, genetically eliminating senescent astrocytes and microglia reduced neuronal tau phosphorylation and NFT deposition. Moreover, the authors found similar effects when they treated tauP301S mice with a 'senolytic' compound, which triggers pharmacological removal of senescent cells. Together, Bussian and colleagues' data clearly demonstrate that tauP301S expression in neurons can induce senescence in brain astrocytes and microglia. In turn, these senescent glia affect the ability of neurons to regulate tau phosphorylation and aggregation (Fig. 1). Whether by releasing signalling molecules that directly affect tau or through the effects of SASP factors (or both), glial senescence ultimately promotes neuronal degeneration.

Bussian and co-workers' findings point to several avenues for future study. First, the signals from tauP301S-expressing neurons that induce senescence in glia should be defined. Similarly, the mechanisms by which senescent astrocytes and microglia signal back to neurons remain to be determined. It will also be interesting to understand whether the same glia-derived signals affect both tau pathology and neuronal survival, and whether astrocytes and microglia send the same or distinct signals. The answers to these questions are likely to have broader implications for understanding neurodegenerative diseases more generally.

Finally, the current study adds to the growing body of evidence indicating that senolytic treatments could benefit people who have a wide range of conditions[1,2]. Of immediate interest is whether removal of senescent cells can decrease disease severity in other animal models of neurodegeneration. The authors removed senescent cells throughout the lives of their animals, but it will also be valuable to determine whether senolytics can have beneficial effects if treatment is started once a disease has progressed to symptomatic stages — a more likely scenario in humans. Finally, it will be crucial to determine whether the processes uncovered in this paper are evolutionarily conserved in humans. If so, perhaps senolytic treatments can benefit people, as promised by this and other mouse studies[1,2]. ■

Jay Penney *and* **Li-Huei Tsai** *are at the Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.*
e-mails: jpenney@mit.edu; lhtsai@mit.edu

1. Childs, B. G. *et al. Nature Rev. Drug Discov.* **16**, 718–735 (2017).
2. Baker, D. J. & Petersen, R. C. *J. Clin. Invest.* **128**, 1208–1216 (2018).
3. Bussian, T. J. *et al. Nature* **562**, 578–582 (2018).
4. Sharpless, N. E. & Sherr, C. J. *Nature Rev. Cancer* **15**, 397–408 (2015).
5. Chinta, S. J. *et al. Cell Rep.* **22**, 930–940 (2018).
6. Bhat, R. *et al. PLoS ONE* **7**, e45069 (2012).
7. Baker, D. J. *et al. Nature* **479**, 232–236 (2011).
8. Baker, D. J. *et al. Nature* **530**,184–189 (2016).
9. Iqbal, K., Liu, F. & Gong, C.-X. *Nature Rev. Neurol.* **12**, 15–27 (2016).

This article was published online on 19 September 2018.

https://doi.org/10.1038/s41586-018-0624-y

Comprehensive measurement of *pp*-chain solar neutrinos

The Borexino Collaboration*

About 99 per cent of solar energy is produced through sequences of nuclear reactions that convert hydrogen into helium, starting from the fusion of two protons (the *pp* chain). The neutrinos emitted by five of these reactions represent a unique probe of the Sun's internal working and, at the same time, offer an intense natural neutrino beam for fundamental physics. Here we report a complete study of the *pp* chain. We measure the neutrino–electron elastic-scattering rates for neutrinos produced by four reactions of the chain: the initial proton–proton fusion, the electron-capture decay of beryllium-7, the three-body proton–electron–proton (*pep*) fusion, here measured with the highest precision so far achieved, and the boron-8 beta decay, measured with the lowest energy threshold. We also set a limit on the neutrino flux produced by the ^{3}He–proton fusion (hep). These measurements provide a direct determination of the relative intensity of the two primary terminations of the *pp* chain (*pp*-I and *pp*-II) and an indication that the temperature profile in the Sun is more compatible with solar models that assume high surface metallicity. We also determine the survival probability of solar electron neutrinos at different energies, thus probing simultaneously and with high precision the neutrino flavour-conversion paradigm, both in vacuum and in matter-dominated regimes.

In 1937, Gamov and von Weizsäcker[1,2] suggested that the Sun is powered by a chain of nuclear reactions initiated by proton–proton fusion and leading to the production of ^{4}He. This idea was further developed by Bethe and Critchfield[3]. At about the same time von Weizsäcker and independently Bethe proposed an alternative mechanism, namely the carbon–nitrogen–oxygen cycle (CNO cycle)[4], a closed-loop chain of nuclear reactions catalysed by ^{12}C, ^{14}N and ^{16}O nuclei in which four protons are converted into ^{4}He. Although the CNO cycle was incorrectly considered to be the main source of energy in the Sun (mainly because of the overestimation of the Sun's central temperature available at that time), the debate on the role of the CNO cycle in the Sun is still relevant today. Indeed, a direct measure of its importance is missing, although theory predicts that it cannot contribute more than about 1% of the solar luminosity. Conversely, it is now understood to be the main source of energy in stars heavier than the Sun. More historical details can be found in ref. [5].

The Sun and lower-mass stars are predominantly powered by the *pp* chain (see Fig. 1), which has been thoroughly studied by Fowler and co-workers in the 1950s[6]. He and A. Cameron also pointed out that the detection of solar neutrinos could be a direct way of testing theoretical solar models. Their intuition was correct and neutrinos are now considered to be the sole direct probes of the Sun's core and of solar energy generation.

Neutrinos are copiously emitted in the primary *pp* fusion reaction of the chain and, to a minor extent, in the alternative three-body *pep* process and in the two secondary branches *pp*-II (^{7}Be neutrinos) and *pp*-III (^{8}B neutrinos). Experimentally, solar neutrinos have been studied since the late 1960s by radiochemical experiments (Homestake[7], SAGE[8] and GALLEX[9]) which, however, could only provide a measurement of their integrated flux on Earth above a threshold. Prior to the establishment of the Borexino project, only ^{8}B neutrinos (<0.01% of the total flux) have been measured individually by KamiokaNDE/SuperKamiokande[10] and the Sudbury Neutrino Observatory (SNO)[11]. Their measurements have definitively proved that neutrinos undergo leptonic flavour (that is, electronic, muonic or tauonic) conversion in the Sun's matter, enhanced through the MSW (Mikheyey–Smirnov–Wolfenstein) mechanism[12–14]. For an historical review of solar-neutrino astronomy and of its impact on solar and neutrino physics see, for example, refs [15–17].

The measurement of all neutrino components is the most direct way to test the standard solar model (SSM)[15] and to validate our theoretical understanding of the properties of the Sun's core. The first theoretical predictions of neutrino fluxes were made in the 1960s by J. Bahcall and his collaborators and have subsequently been refined by many theoretical groups[18]. Despite the results delivered by solar-neutrino experiments, important questions about the Sun remain unanswered. For example, the solar metallicity, that is, the abundance of elements heavier than He, is poorly understood, even though it is a fundamental parameter for the determination of the physical properties of the Sun. A precise measurement of the solar neutrino fluxes comprising the *pp* chain and the CNO cycle would directly settle the controversy between high-metallicity (HZ) and low-metallicity (LZ) SSMs[18] (see Methods). This study is a step in this direction.

Solar neutrinos are also powerful probes of neutrino properties. First, they allow the determination of oscillation parameters, especially the θ_{12} mixing angle and, to a lesser degree, the Δm^2_{12} mass splitting. Second, the measurement of the electron neutrino survival probability (P_{ee}) as a function of neutrino energy allows us to probe directly the MSW-LMA mechanism of neutrino oscillations[19] and to search for deviations that could indicate the presence of physics beyond the standard model.

Running continuously since 2007, the Borexino experiment has measured, one after another, ^{7}Be neutrinos[20–22], *pep* neutrinos[23], ^{8}B neutrinos[24] and *pp* neutrinos[25]. Here we report the simultaneous precision spectroscopic measurement of the complete *pp* chain and its implications for both solar and neutrino physics.

Borexino and the solar-neutrino analysis

Borexino is a liquid-scintillator experiment at the Laboratori Nazionali del Gran Sasso in Italy[26]. Given the tiny cross-section of neutrino interactions with electrons ($\sigma \approx 10^{-44}$ cm^2 to 10^{-45} cm^2 for the solar-neutrino energy range), the rates expected in Borexino are small,

*A list of participants and their affiliations appears at the end of the paper.

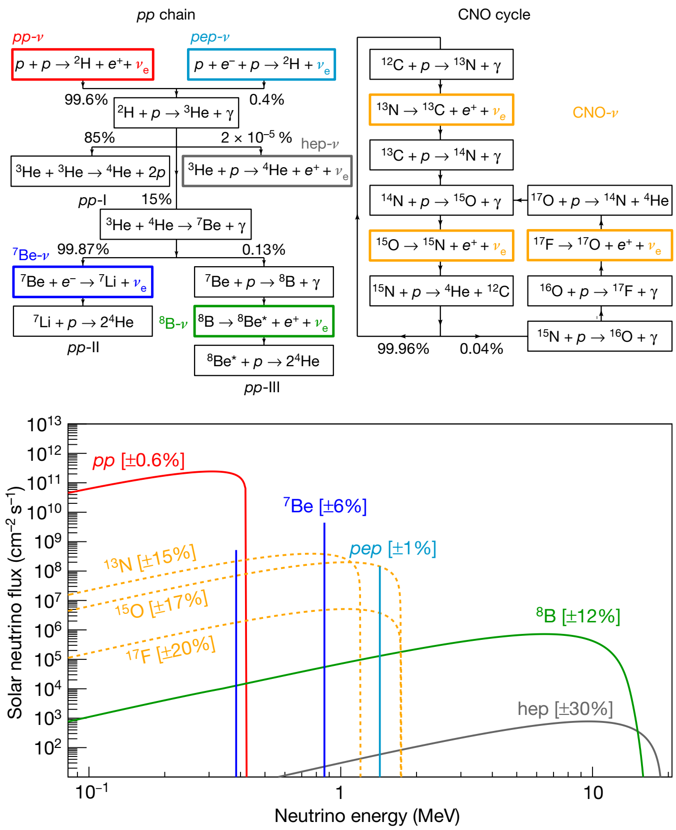


Fig. 1 | Nuclear fusion sequences and neutrino energy spectrum. Schematic view of the *pp* and CNO nuclear fusion sequences. The solar-neutrino energy spectrum is obtained from http://www.sns.ias.edu/~jnb/, using the updated fluxes taken from ref. [18]. The flux (vertical scale) is given in units of $cm^{-2}\,s^{-1}\,MeV^{-1}$ for continuum sources and in $cm^{-2}\,s^{-1}$ for monoenergetic sources.

ranging from less than one to a few tens of counts per day per 100 tons (t) for different solar-neutrino components. To cope with such a low event rate, Borexino has a large target mass (about 300 t) and is housed deep underground, under 3,800 m water equivalent of dolomitic rock that suppresses the flux of cosmic radiation by a factor of approximately one million. For more details on the detector, see Methods.

Radioactive decays of unstable isotopes contained in the scintillator or in the materials surrounding it represent the main sources of background (referred to as internal and external, respectively). Whereas external background is greatly reduced by concentric layers of high-purity materials surrounding the scintillator and by the selection of a centrally located software-defined fiducial volume, most of the internal background can only be cut down by means of liquid-scintillator purification. Particularly, interactions of beta particles (β; electrons and positrons) and of gamma particles (γ; high-energy photons) must be reduced to very low levels, since they cannot be distinguished from neutrino interactions on an event-by-event basis. Borexino has reached unprecedented levels of scintillator radio-purity. As an example, one gram of liquid scintillator contains less than 9.4×10^{-20} grams of uranium-238 and less than 5.7×10^{-19} grams of thorium-232 (95% confidence level, C.L.), a concentration about ten orders of magnitude smaller than in any natural material on Earth. This low level of background has enabled real-time detection of solar neutrinos with an energy threshold of 0.19 MeV, and allowed us to perform the complete spectroscopy of the *pp* chain.

Solar neutrinos reach the Earth as a mixture of all neutrino flavours (electronic, muonic, and tauonic) owing to the flavour-conversion mechanism enhanced by the MSW effect (see Methods). Borexino detects them by means of their weak elastic scattering off electrons. A fraction of the incoming neutrino energy E_ν is transferred to one electron, which deposits it in the liquid scintillator. The scintillator light is detected by about 2,000 photomultiplier tubes, which ensure high detection efficiency of photoelectrons produced by incident optical photons at their photocathodes. For ^{7}Be ($E_\nu = 0.384$ MeV and 0.862 MeV) and *pep* ($E_\nu = 1.44$ MeV) neutrinos, the induced electron recoil endpoints are 0.230 MeV, 0.665 MeV and 1.22 MeV, respectively. For the continuous *pp* and ^{8}B spectra, they are 0.261 MeV and 15.2 MeV, respectively.

The detected light and its time distribution among photomultiplier tubes yield three important quantities for each interaction event in the detector: its deposited energy, roughly proportional to the total number of detected photoelectrons; its position within the detector, obtained from the analysis of the photon arrival times at each photomultiplier tube; and its particle identification, based on a pulse-shape discrimination method that exploits the different time structure of liquid-scintillator light pulses produced by different particles (electrons, positrons, α particles and protons)[27]. For reference, a 1-MeV electron produces on average 500 photoelectrons in 2,000 photomultiplier tubes, its energy is measured with $\sigma \approx 50$ keV and its position is reconstructed[28,29] with $\sigma \approx 12$ cm.

We divided the analysis into two energy regions that are affected by different backgrounds, which need to be handled differently: a low-energy region (LER) of 0.19–2.93 MeV, to measure the *pp*, ^{7}Be and *pep* neutrino interaction rates, and a high-energy region (HER) of 3.2–16 MeV, to measure ^{8}B neutrinos. For the same reason, the HER is further divided into two subregions, below and above 5.7 MeV (HER-I and HER-II). The measurement of ^{8}B neutrinos cannot be extended below 3.2 MeV because of the 2.614-MeV γ-ray background from ^{208}Tl decays, originating from trace ^{232}Th contamination of the thin nylon liquid-scintillator containment vessel.

The reconstructed position of each event within the detector allows us to define a fiducial volume optimized differently for the analysis in the LER and HER-I/II. The LER fiducial volume is chosen to suppress external γ-rays from ^{40}K, ^{214}Bi and ^{208}Tl contained in materials surrounding the scintillator and consists of the innermost 71.3 t of scintillator selected with a radial cut (radius $R < 2.8$ m) and a cut in the vertical direction (-1.8 m $< z < 2.2$ m). The HER is above the energy of the aforementioned γ-rays. The analysis in HER-I requires only a $z < 2.5$ m cut to suppress background events related to a small pinhole in the inner vessel that causes liquid scintillator to leak into the region outside the inner vessel. The total selected mass in this case is 227.8 t. In contrast, the analysis in HER-II uses the entire scintillator volume, 266 t, since the above-mentioned background does not affect this energy window.

The LER analysis uses exclusively Borexino Phase-II data collected between December 2011 and May 2016, in which the internal ^{85}Kr and ^{210}Bi contamination was reduced with respect to Borexino Phase-I, thanks to a liquid-scintillator purification campaign carried out in 2010 and 2011. The total LER exposure is 1,291.51 days $\times$ 71.3 t. With the exception of ^{208}Tl decays (Q-value, total energy released in the decay, about 5 MeV), the HER is above the natural, long-lived radioactive background, making it possible to use a larger dataset, collected between January 2008 and December 2016, for a total exposure of 2,062.4 days $\times$ 227.8 (266.0) t for HER-I (or HER-II), respectively.

The analysis proceeds in two steps: (1) the event selection, with a different set of cuts in the three energy regions to maximize the signal-to-background ratio, and (2) the extraction of the neutrino and residual background rates with a combined fit of distributions of global quantities built for the events surviving the cuts. The main event selection criteria are conceptually similar for the LER and the HER and are conceived to: reject cosmic muons surviving the mountain shield[30]; reduce the cosmogenic background (that is, radioactive elements produced in muon-induced nuclear spallation processes); and select an optimal spatial region of the scintillator (the fiducial volume). More details on the cuts are discussed in Methods.

Several backgrounds, listed in Table 1 and described in detail in Methods, survive the event selection cuts. To disentangle the neutrino signal from these backgrounds, two different fitting strategies are adopted for the LER and the HER. The LER analysis follows a

Table 1 | Rates of residual backgrounds

Background LER	Rate (Bq per 100 t)
^{14}C(0.156 MeV, β^-)	[40.0 ± 2.0]
Background LER	**Rate (counts per day per 100 t)**
^{85}Kr (0.687 MeV, β^-) (internal)	6.8 ± 1.8
^{210}Bi (1.16 MeV, β^-) (internal)	17.5 ± 1.9
^{11}C (1.02–1.98 MeV, β^+) (internal)	26.8 ± 0.2
^{210}Po (5.3 MeV, α) (internal)	260.0 ± 3.0
^{40}K (1.460 MeV, γ) (external)	1.0 ± 0.6
^{214}Bi (<1.764 MeV, γ) (external)	1.9 ± 0.3
^{208}Tl (2.614 MeV, γ) (external)	3.3 ± 0.1
Background HER-I	**Rate (counts per day per 227.8 t)**
μ, cosmogenics, ^{214}Bi (internal)	[$6.1^{+8.7}_{-3.1} \times 10^{-3}$]
(α, n) (external)	0.224 ± 0.078
^{208}Tl(5.0 MeV, β^-, γ) (internal)	[0.042 ± 0.008]
^{208}Tl(5.0 MeV, β^-, γ) (emanated)	0.469 ± 0.063
^{208}Tl(5.0 MeV, β^-, γ) (surface)	1.090 ± 0.046
Background HER-II	**Rate (counts per day per 266.0 t)**
μ, cosmogenics (internal)	[$3.8^{+14.6}_{-0.1} \times 10^{-3}$]
(α, n) (external)	0.239 ± 0.022

Residual background is due to β^- (electrons), β^+ (positrons), γ (gammas), μ (muons), α (alpha particles) and n (neutrons). The background rates are obtained by the fit to the energy spectrum of collected events in the three energy regions used in this study (LER, HER-I and HER-II). We report in parentheses the Q-value and type of particle for each background. The rates in square brackets are estimated independently and are constrained in the fit. Background can be internal (that is, due to events uniformly distributed in the scintillator volume) or external (that is, due to events from sources surrounding the scintillator).

multivariate approach, simultaneously fitting the energy spectrum, the spatial and the pulse-shape estimator distributions. In the HER-I and HER-II, a fit of the radial distribution of events is performed to separate the ^{8}B neutrino signal (uniformly distributed in the scintillator) from the external background.

Some residual background rates are measured independently, whenever possible, and are constrained in the fit (values between squared brackets in Table 1). The remaining background rates are left free to vary and are returned by the fit together with the neutrino rates.

The results of the fit are exemplified in Fig. 2a, which shows the energy spectrum in the LER after applying the threefold coincidence method (TFC) to reduce the ^{11}C cosmogenic background (see Methods); Fig. 2b shows the radial distribution of the events in HER-I. The different contributions from signal and background as determined by the fit are superimposed to data in the plots. The results of the fit for the untagged backgrounds are summarized in Table 1.

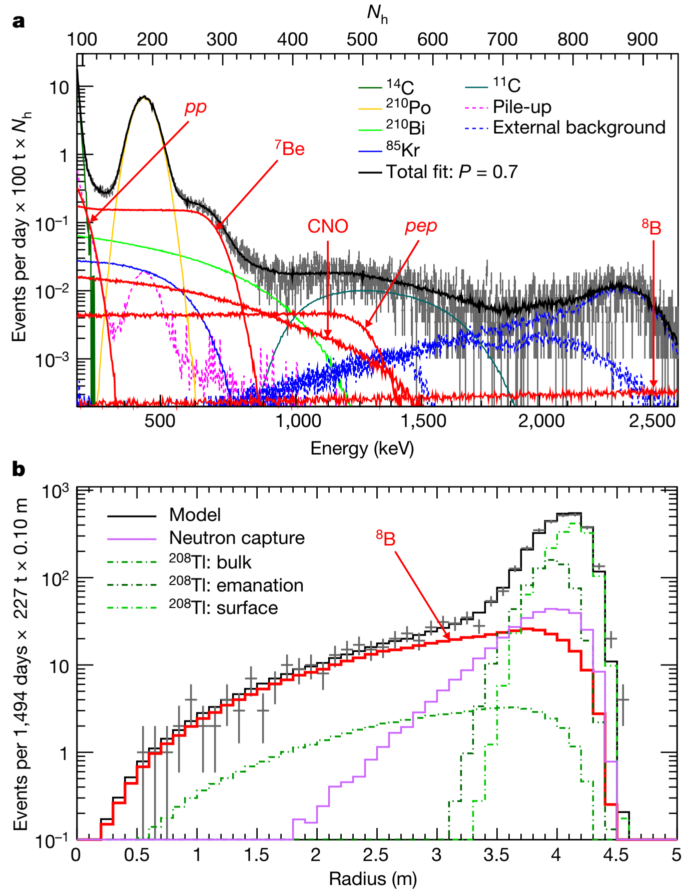


Fig. 2 | Results of the fit used to extract the neutrino signal. Distributions of events after selection cuts and corresponding fits with neutrino and background components. **a**, TFC-subtracted energy spectrum with suppressed ^{11}C background in LER. The horizontal upper scale is in units of N_h, that is, the total number of photons collected for each event. **b**, Radial distribution of events in HER-I.

Results

The high-precision solar-neutrino results obtained in this work are summarized in Table 2. The second column reports the measured rates. In the third column, we translate these measurements into the corresponding solar-neutrino fluxes using the known electron and μ/τ neutrino cross-sections[27] and the flavour composition calculated according to the MSW-LMA paradigm (mass and mixing parameters from ref. [19]). The fourth column shows the theoretical fluxes predicted by the SSM under the HZ and LZ assumptions[18].

In the LER multivariate fit, performed to extract the *pp*, *pep* and ^{7}Be neutrino rates, we first constrain the CNO neutrino interaction rate to the value predicted by the HZ-SSM assuming the MSW-LMA scenario (4.92 ± 0.55 counts per day per 100 t)[18,19], then, separately, to the LZ-SSM predictions (3.52 ± 0.37 counts per day per 100 t). Only the *pep* neutrino rate is slightly influenced by this constraint and thus two results for it are reported. In both cases, the absence of the *pep* reaction in the Sun is rejected with $>5\sigma$ significance, enough to definitively claim discovery of solar *pep* neutrinos. The contribution of ^{8}B neutrinos in the LER is very small and its rate was constrained to the value obtained from the HER analysis. Statistical uncertainties are evaluated by profiling the likelihood using Wilks's approximation, whose adequacy in this case is confirmed by Monte Carlo simulations. The ^{7}Be solar-neutrino flux is determined with a total uncertainty of 2.7%, a factor of 1.8 improvement with respect to our previous result[22] and a factor of two smaller than the theoretical uncertainty. The *pp* interaction rate is consistent with our previous result[25] and has an uncertainty of 9.5%. Fits were performed with several hundred configurations, yielding results whose spread is incorporated in the systematic uncertainties (see Methods for more details).

The ^{8}B solar-neutrino flux derived from our measured rate in the entire HER is $(5.68^{+0.39}_{-0.41}{}^{+0.03}_{-0.03}) \times 10^6\ \mathrm{cm^{-2}\,s^{-1}}$, consistent with our previous result[24] and with the high-precision determination by SuperKamiokande[31] and SNO[32]. The equivalent-flavour stable ^{8}B flux, that is, the flux obtained attributing the measured rate entirely to electron neutrinos, is $(2.57^{+0.17}_{-0.18}{}^{+0.07}_{-0.07}) \times 10^6\ \mathrm{cm^{-2}\,s^{-1}}$. The uncertainty in the ^{8}B rate determination is 8%, a more than twofold improvement on our previous measurement[24].

The similarity between the electron recoil spectrum induced by CNO neutrinos and the ^{210}Bi β-decay spectrum makes it impossible to disentangle the two contributions with the spectral fit. For this reason, we only provide an upper limit on the CNO neutrino interaction rate. To do so, we also place an indirect constraint on *pep* neutrinos by exploiting the theoretically well known *pp* and *pep* flux ratio. Using values predicted by the HZ-SSM[18] and including the effect of MSW-LMA oscillations[19], the ratio of *pp* and *pep* neutrino interaction rates is 47.8 ± 0.8. Using the ratio predicted by the LZ-SSM, 47.5 ± 0.8, yields

Table 2 | Borexino experimental solar-neutrino results

Solar neutrino	Rate (counts per day per 100 t)	Flux (cm^{-2} s^{-1})	Flux–SSM predictions (cm^{-2} s^{-1})
pp	$134 \pm 10^{+6}_{-10}$	$(6.1 \pm 0.5^{+0.3}_{-0.5}) \times 10^{10}$	$5.98(1.0 \pm 0.006) \times 10^{10}$ (HZ) $6.03(1.0 \pm 0.005) \times 10^{10}$ (LZ)
^{7}Be	$48.3 \pm 1.1^{+0.4}_{-0.7}$	$(4.99 \pm 0.11^{+0.06}_{-0.08}) \times 10^9$	$4.93(1.0 \pm 0.06) \times 10^9$ (HZ) $4.50(1.0 \pm 0.06) \times 10^9$ (LZ)
pep (HZ)	$2.43 \pm 0.36^{+0.15}_{-0.22}$	$(1.27 \pm 0.19^{+0.08}_{-0.12}) \times 10^8$	$1.44(1.0 \pm 0.01) \times 10^8$ (HZ) $1.46(1.0 \pm 0.009) \times 10^8$ (LZ)
pep (LZ)	$2.65 \pm 0.36^{+0.15}_{-0.24}$	$(1.39 \pm 0.19^{+0.08}_{-0.13}) \times 10^8$	$1.44(1.0 \pm 0.01) \times 10^8$ (HZ) $1.46(1.0 \pm 0.009) \times 10^8$ (LZ)
$^8B_{HER-I}$	$0.136^{+0.013+0.003}_{-0.013-0.003}$	$(5.77^{+0.56+0.15}_{-0.56-0.15}) \times 10^6$	$5.46(1.0 \pm 0.12) \times 10^6$ (HZ) $4.50(1.0 \pm 0.12) \times 10^6$ (LZ)
$^8B_{HER-II}$	$0.087^{+0.080+0.005}_{-0.010-0.005}$	$(5.56^{+0.52+0.33}_{-0.64-0.33}) \times 10^6$	$5.46(1.0 \pm 0.12) \times 10^6$ (HZ) $4.50(1.0 \pm 0.12) \times 10^6$ (LZ)
$^8B_{HER}$	$0.223^{+0.015+0.006}_{-0.016-0.006}$	$(5.68^{+0.39+0.03}_{-0.41-0.03}) \times 10^6$	$5.46(1.0 \pm 0.12) \times 10^6$ (HZ) $4.50(1.0 \pm 0.12) \times 10^6$ (LZ)
CNO	<8.1 (95% C.L.)	$<7.9 \times 10^8$ (95% C.L.)	$4.88(1.0 \pm 0.11) \times 10^8$ (HZ) $3.51(1.0 \pm 0.10) \times 10^8$ (LZ)
hep	<0.002 (90% C.L.)	$<2.2 \times 10^5$ (90% C.L.)	$7.98(1.0 \pm 0.30) \times 10^3$ (HZ) $8.25(1.0 \pm 0.12) \times 10^3$ (LZ)

Measured neutrino rates (second column): for *pp*, ^{7}Be, *pep* and CNO neutrinos we quote the total counts without any threshold; for ^{8}B and hep neutrinos we quote the counts above the corresponding analysis threshold. Neutrino fluxes (third column) are obtained from the measured rates assuming the MSW-LMA oscillation parameters[19], standard neutrino–electron cross-sections[27] and a density of electrons in the scintillator of $(3.307 \pm 0.003) \times 10^{31}$ electrons per 100 t. All fluxes are integral values without any threshold. The result for *pep* neutrinos depends on whether we assume HZ or LZ SSM predictions to constrain the CNO neutrino flux. The last column shows the fluxes predicted by the SSM for the HZ or LZ hypotheses[18].

identical results. We obtain an upper limit of <8.1 counts per day per 100 t (95% C.L.) for the CNO neutrino interaction rate, in agreement with the Borexino sensitivity to CNO studied with Monte Carlo.

For completeness, we also perform a search for the hep neutrinos, emitted by the proton capture reaction of ^{3}He (Fig. 1). The expected flux is more than two orders of magnitude smaller than that of ^{8}B neutrinos. Despite their higher end-point energy, this signal in Borexino is extremely small and covered by background, particularly cosmogenic ^{11}Be decays ($Q = 11.5$ MeV, β^-, $\tau = 19.9$ s) and ^{8}B neutrinos. We perform a dedicated analysis on the whole dataset (0.8 kt yr) and in the energy region 11–20 MeV we find 10 ± 3 events, consistent with the expected background. We obtain an upper limit for the hep neutrino flux of 2.2×10^5 cm^{-2} s^{-1} (90% C.L.) to be compared with the expected flux 7.98×10^3 cm^{-2} s^{-1} (8.25×10^3 cm^{-2} s^{-1}) assuming the HZ (LZ) SSM.

Discussion and outlook

The measurements reported in this work represent a complete study of the solar *pp* chain and of its different terminations by means of neutrino detection in a single detector and with a uniform data analysis procedure. These measurements can be used either to test the MSW-LMA paradigm assuming SSM flux predictions or, alternatively, to probe our understanding of solar physics assuming the validity of the neutrino oscillation mechanism.

The interaction rates of *pp*, ^{7}Be, *pep* and ^{8}B neutrinos reported in Table 2 can be used to infer the electron neutrino survival probability at different energies. Assuming the HZ-SSM fluxes[18] and standard neutrino-electron cross-sections[27], we obtain the electron neutrino survival probabilities for each solar-neutrino component: $P_{ee}(pp, 0.267\ \text{MeV}) = 0.57 \pm 0.09$, $P_{ee}(^7\text{Be}, 0.862\ \text{MeV}) = 0.53 \pm 0.05$, and $P_{ee}(pep, 1.44\ \text{MeV}) = 0.43 \pm 0.11$. The quoted errors include the uncertainties on the SSM solar-neutrino flux predictions. The ^{8}B electron neutrino survival probability is calculated in each HER range following the procedure described in ref. [24]. We obtain $P_{ee}(^8\text{B}_{\text{HER}}, 8.1\ \text{MeV}) = 0.37 \pm 0.08$, $P_{ee}(^8\text{B}_{\text{HER-I}}, 7.4\ \text{MeV}) = 0.39 \pm 0.09$, and $P_{ee}(^8\text{B}_{\text{HER-II}}, 9.7\ \text{MeV}) = 0.35 \pm 0.09$. These results are summarized in Fig. 3. For non-monoenergetic components, that is, *pp* and ^{8}B neutrinos, the P_{ee} value is quoted for the average energy of neutrinos that produce scattered electrons in the given energy range.

Borexino provides the most precise measurement of the P_{ee} in the LER, where flavour conversion is vacuum-dominated. At higher energy, where flavour conversion is dominated by matter effects in the Sun, the Borexino results are in agreement with the high-precision measurements performed by SuperKamiokande[31] and SNO[32]. Borexino is the only experiment that can simultaneously test neutrino flavour conversion both in the vacuum and in the matter-dominated regime. We performed a likelihood ratio test to compare our data with the MSW-LMA and the vacuum-LMA predictions (pink and grey bands in Fig. 3, respectively). Our data disfavour the vacuum-LMA hypothesis at 98.2% C.L. (see Methods). Overall, the results are in excellent agreement with the expectations from the MSW-LMA paradigm with the oscillation parameters indicated in ref. [19].

Since solar neutrinos are detected on Earth only about 8 min after being produced, they provide a real-time picture of the core of the Sun. In particular, the neutrino fluxes determined experimentally can be used to derive the total power generated by nuclear reactions in the Sun's core[33]. By using exclusively the new Borexino results reported in

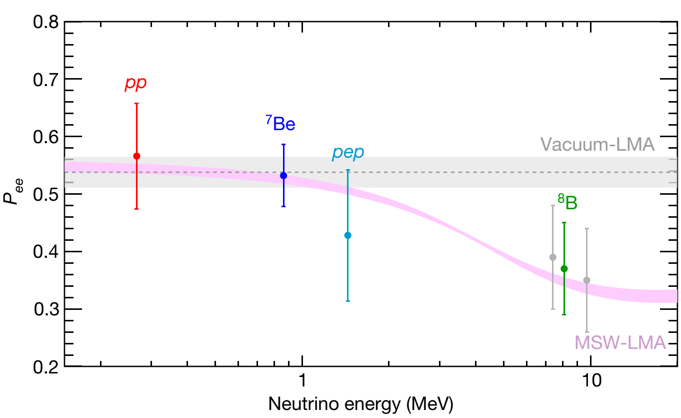


Fig. 3 | Electron neutrino survival probability P_{ee} as a function of neutrino energy. The pink band is the $\pm 1\sigma$ prediction of MSW-LMA with oscillation parameters determined from ref. [19]. The grey band is the vacuum-LMA case with oscillation parameters determined from refs [38,39]. Data points represent the Borexino results for *pp* (red), ^{7}Be (blue), *pep* (cyan) and ^{8}B (green for the HER range, and grey for the separate HER-I and HER-II sub-ranges), assuming HZ-SSM. ^{8}B and *pp* data points are set at the mean energy of neutrinos that produce scattered electrons above the detection threshold. The error bars include experimental and theoretical uncertainties.

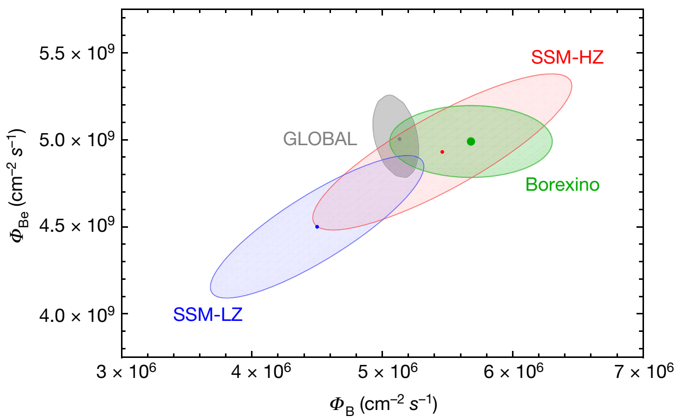


Fig. 4 | Borexino results and analysis in the $\Phi(^7\mathrm{Be})$–$\Phi(^8\mathrm{B})$ space. Borexino results for ^{7}Be and ^{8}B neutrino fluxes (green point and shaded area). Allowed contours in the $\Phi(^7\mathrm{Be})$–$\Phi(^8\mathrm{B})$ space are obtained by combining these new results with all solar and KamLAND data in a global analysis, and leaving free the oscillation parameters θ_{12} and Δm^2_{12} (grey ellipse, marked as GLOBAL). The theoretical prediction for the low-metallicity (LZ) (blue) and the high-metallicity (HZ) (red) Standard Solar Models (SSM)[18] are also shown. The fit returns the following oscillation parameters: $\tan^2\theta_{12} = 0.47 \pm 0.03$ and $\Delta m^2_{12} = (7.5 \times 10^{-5}) \pm 0.03$, in agreement with what is reported in ref. [19] ($\sin^2\theta_{13}$ is fixed to 0.0217; ref. [19]). All contours correspond to 68.27% C.L.

Table 2, we find $L = (3.89^{+0.35}_{-0.42}) \times 10^{33}$ erg s^{-1}, in agreement with the luminosity calculated using the well measured photon output[34,35], $L = (3.846 \pm 0.015) \times 10^{33}$ erg s^{-1}. This confirms experimentally the nuclear origin of the solar power with the best precision obtained by a single solar-neutrino experiment. Considering that it takes around 10^5 years for radiation to flow from the energy-producing region to the surface of the Sun, this comparison proves also that the Sun has been in thermodynamic equilibrium over this timescale.

Furthermore, we derive the ratio $R_{\mathrm{I/II}}$ between the ^{3}He–^{4}He and the ^{3}He–^{3}He fusion rates, which quantifies the relative intensity of the two primary terminations of the *pp* chain (*pp*-II and *pp*-I; see Fig. 1), a critical probe of solar fusion. Neglecting the ^{8}B neutrino contribution, this ratio can be extracted from the measured *pp* and ^{7}Be neutrino fluxes by the relation[36], $R_{\mathrm{I/II}} = 2\Phi(^7\mathrm{Be})/[\Phi(pp) - \Phi(^7\mathrm{Be})]$. We find $R_{\mathrm{I/II}} = 0.178^{+0.027}_{-0.023}$, in agreement with the most up-to-date predicted values of $R_{\mathrm{I/II}} = 0.180 \pm 0.011$ (HZ) and 0.161 ± 0.010 (LZ)[18].

Finally, the Borexino measurements can be used to test the predictions of SSMs with different metallicity. Indeed, the assumed metallicity determines the opacity of solar plasma and, as a consequence, regulates the central temperature of the Sun and the branching ratios of the different *pp*-chain terminations. To perform this test, we use only the results for ^{7}Be and ^{8}B neutrinos, whose fluxes are very different in the HZ- and the LZ-SSM theoretical predictions (differences of 9% and 18%, respectively). Figure 4 shows the results of Borexino (green-shaded ellipse), together with the predictions for the HZ- and LZ-SSMs[18] (red- and blue-shaded ellipses, respectively). Note that the errors in the Borexino measurements are in both cases smaller than the theoretical uncertainties. The theoretical error budget is dominated by uncertainties on the astrophysical factor S_{34} of the $^3\mathrm{He} + {}^4\mathrm{He}$ reaction, on the opacity of the Sun, and on the astrophysical factor S_{17} of the $p + {}^7\mathrm{Be}$ reaction as discussed in ref. [18].

The Borexino results are compatible with the temperature profiles predicted by both HZ- and LZ-SSMs. However, the ^{7}Be and ^{8}B solar-neutrino fluxes measured by Borexino provide an interesting hint in favour of the HZ-SSM prediction. A frequentist hypothesis test based on a likelihood-ratio test statistics (HZ versus LZ) was performed by computing the probability distribution functions with a Monte Carlo approach. Assuming HZ to be true, our data disfavour LZ at 96.6% C.L. This constraint is slightly stronger than our sensitivity (the median sensitivity is at 94.2% C.L.). A Bayesian hypothesis test[37] yields a Bayes factor of 4.9, confirming a mild preference for HZ (see Methods for more details on both the frequentist and Bayesian studies).

For the sake of completeness, we performed a global fit including the results presented in this work together with all the other solar + KamLAND data. Following the procedure described in ref. [27], we leave the oscillation parameters θ_{12}, Δm^2_{12} and the ^{7}Be and ^{8}B neutrino fluxes free to vary in the fit. Figure 4 shows the allowed regions in the $\Phi(^7\mathrm{Be})$–$\Phi(^8\mathrm{B})$ space determined from this global analysis. The oscillation parameters returned by the fit are consistent with the ones obtained in ref. [19]. It is clear from the output of this global fit that when the Borexino results are combined with those of all other solar-neutrino experiments, the small hint towards HZ further weakens.

In summary, we have reported simultaneous measurements of solar neutrinos from all the reactions belonging to the *pp* nuclear fusion chain. This study confirms the nuclear origin of the solar power and provides the most complete real-time insight into the core of our Sun so far.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0624-y.

Received: 23 March; Accepted: 23 August 2018;
Published online 24 October 2018.

1. Atkinson, R. & Houtermans, F. Zur Frage der Aufbaumöglichkeit der Elemente in Sternen. *Z. Phys.* **54**, 656 (1929).
2. von Weizsäcker, C. F. Über Elementumwandlungen im Innern der Sterne I. *Phys. Z.* **38**, 176 (1937).
3. Bethe, H. A. & Critchfield, C. L. The formation of deuterons by proton combination. *Phys. Rev.* **54**, 248 (1938).
4. Bethe, H. Energy production in stars. *Phys. Rev.* **55**, 434 (1939).
5. Bahcall, J. N. *How the Sun Shines*. https://www.nobelprize.org/nobel_prizes/themes/physics/fusion/ (Nobel Media, Stockholm, 2000).
6. Fowler, W. Experimental and theoretical nuclear astrophysics; the quest for the origin of the elements: Nobel prize lecture. *Rev. Mod. Phys.* **56**, 149 (1984).
7. Davis, R. Nobel lecture: a half-century with solar neutrinos. *Rev. Mod. Phys.* **75**, 985 (2003).
8. Abdurashitov, J. et al. Results from SAGE (the Russian-American gallium solar neutrino experiment). *Phys. Lett. B* **328**, 234 (1994).
9. Anselmann, P. et al. Solar neutrinos observed by GALLEX at Gran Sasso. *Phys. Lett. B* **285**, 376 (1992).
10. Hirata, K. et al. Observation of ^{8}B solar neutrinos in the Kamiokande-II detector. *Phys. Rev. Lett.* **63**, 16 (1989).
11. Ahmad, Q. et al. Direct evidence for neutrino flavor transformation from neutral-current interactions in the Sudbury Neutrino Observatory. *Phys. Rev. Lett.* **89**, 011301 (2002).
12. Pontecorvo, B. Neutrino experiments and the problem of conservation of leptonic charge. *Zh. Eksp. Teor. Fiz.* **53**, 1717 (1967).
13. Wolfenstein, L. Neutrino oscillations in matter. *Phys. Rev. D* **17**, 2369 (1978).
14. Mikheyev, S. & Smirnov, A. Resonant amplification of neutrino oscillations in matter and spectroscopy of solar neutrinos. *Sov. J. Nucl. Phys.* **42**, 913 (1985).
15. Bahcall, J. & Davis, R. The evolution of neutrino astronomy. *Publ. Astron. Soc. Pacif.* **112**, 429 (2000).
16. Haxton, W., Hamish Robertson, R. & Serenelli, A. Solar neutrinos: status and prospects. *Annu. Rev. Astron. Astrophys.* **51**, 21 (2013).
17. Bahcall, J. N. *Neutrino Astrophysics* (Cambridge Univ. Press, Cambridge, 1989).
18. Vinyoles, N. et al. A new generation of standard solar models. *Astrophys. J.* **835**, 202 (2017).
19. Esteban, I. et al. Updated fit to three neutrino mixing: exploring the accelerator-reactor complementarity. *J. High Energy Phys.* **1701**, 087 (2017).
20. Arpesella, C. et al. First real time detection of ^{7}Be solar neutrinos by Borexino. *Phys. Lett. B* **658**, 101 (2008).
21. Arpesella, C. et al. Direct measurement of the ^{7}Be solar neutrino flux with 192 days of Borexino data. *Phys. Rev. Lett.* **101**, 091302 (2008).
22. Bellini, G. et al. Precision measurement of the ^{7}Be solar neutrino interaction rate in Borexino. *Phys. Rev. Lett.* **107**, 141302 (2011).
23. Bellini, G. et al. First evidence of *pep* solar neutrinos by direct detection in Borexino. *Phys. Rev. Lett.* **108**, 051302 (2012).
24. Bellini, G. et al. Measurement of the solar ^{8}B neutrino rate with a liquid scintillator target and 3 MeV energy threshold in the Borexino detector. *Phys. Rev. D* **82**, 033006 (2010).
25. Borexino Collaboration. Neutrinos from the primary proton-proton fusion process in the Sun. *Nature* **512**, 383 (2014).
26. Alimonti, G. et al. The Borexino detector at the Laboratori Nazionali del Gran Sasso. *Nucl. Instrum. Meth. A* **600**, 568 (2009).
27. Bellini, G. et al. Final results of Borexino Phase I on low-energy solar neutrino spectroscopy. *Phys. Rev. D* **89**, 112007 (2014).
28. Back, H. et al. Borexino calibrations: hardware, methods and results. *J. Instrum.* **7**, P10018 (2012).

29. Agostini, M. et al. The Monte Carlo simulation of the Borexino detector. *Astropart. Phys.* **97**, 136 (2018).
30. Bellini, G. et al. Muon and cosmogenic neutron detection in Borexino. *J. Instrum.* **6**, P05005 (2012).
31. Abe, K. et al. Solar neutrino measurements in Super-Kamiokande-IV. *Phys. Rev. D* **94**, 052010 (2016).
32. Aharmim, B. et al. Combined analysis of all three phases of solar neutrino data from the Sudbury Neutrino Observatory. *Phys. Rev. C* **88**, 025501 (2013).
33. Bergström, J. et al. Updated determination of the solar neutrino fluxes from solar neutrino data. *J. High Energy Phys.* **2016**, 132 (2016).
34. Chapman, G. A. in *Encyclopedia of Planetary Science and Encyclopedia of Earth Science* 748 (Springer, 1997).
35. Fröhlich, C. & Lean, J. The Sun's total irradiance: cycles, trends and related climate change uncertainties since 1976. *Geophys. Res. Lett.* **25**, 4377 (1998).
36. Bahcall, J. & Pena-Garay, C. A road map to solar neutrino fluxes, neutrino oscillation parameters and tests for new physics. *J. High Energy Phys.* **2003**, 4 (2003).
37. Caldwell, A., Kollar, D., Kroninger, K. BAT—the Bayesian Analysis Toolkit. *Comput. Phys. Commun.* **180**, 2197 (2009).
38. Feng Pen An et al. Measurement of electron antineutrinon oscillation based on 1230 days of operation of the Daya Bay experiment. *Phys. Rev. D* **95**, 072006 (2017).
39. Gando, A. et al. Reactor on-off antineutrino measurement with KamLAND. *Phys. Rev. D* **88**, 033001 (2013).

Acknowledgements The Borexino programme is made possible by funding from INFN (Italy), NSF (USA), BMBF, DFG, HGF and MPG (Germany), RFBR (grants 16-29-13014ofi-m and 17-02-00305A), RSF (grant 17-12-01009) (Russia), and NCN (grant number UMO 2017/26/M/ST2/00915) (Poland). We acknowledge also the computing services of the Bologna INFN-CNAF data centre and LNGS Computing and Network Service (Italy), of Jülich Supercomputing Centre at FZJ (Germany), and of ACK Cyfronet AGH Cracow (Poland). We acknowledge the hospitality and support of the Laboratori Nazionali del Gran Sasso (Italy).

Reviewer information *Nature* thanks A. Serenelli and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions The Borexino detector was designed, constructed, and commissioned by the Borexino Collaboration over the span of more than 15 years. The Borexino Collaboration sets the science goals. Scintillator purification and handling, source calibration campaigns, photomultiplier tube and electronics operations, signal processing and data acquisition, Monte Carlo simulations of the detector, and data analyses were performed by Borexino members, who also discussed and approved the scientific results. This manuscript was prepared by a subgroup of authors appointed by the Collaboration and subjected to an internal collaboration-wide review process. All authors reviewed and approved the final version of the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0624-y.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to the Borexino Collaboration.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The Borexino Collaboration

M. Agostini[1], K. Altenmüller[1], S. Appel[1], V. Atroshchenko[2], Z. Bagdasarian[3], D. Basilico[4], G. Bellini[4], J. Benziger[5], D. Bick[6], G. Bonfini[7], D. Bravo[4,29], B. Caccianiga[4*], F. Calaprice[8], A. Caminata[9], S. Caprioli[4], M. Carlini[7], P. Cavalcante[7,10], A. Chepurnov[11], K. Choi[12], L. Collica[4], D. D'Angelo[4], S. Davini[9], A. Derbin[13], X. F. Ding[7,14], A. Di Ludovico[8], L. Di Noto[9], I. Drachnev[13], K. Fomenko[15], A. Formozov[4,11,15], D. Franco[16], F. Gabriele[7], C. Galbiati[8,14], C. Ghiano[7], M. Giammarchi[4], A. Goretti[7], M. Gromov[11], D. Guffanti[7,14], C. Hagner[6], T. Houdy[16], E. Hungerford[17], Aldo Ianni[7,18], Andrea Ianni[8], A. Jany[19], D. Jeschke[1], V. Kobychev[20], D. Korablev[15], G. Korga[17], D. Kryn[16], M. Laubenstein[7], E. Litvinovich[2,21], F. Lombardi[7,30], P. Lombardi[4], L. Ludhova[3,22], G. Lukyanchenko[2], L. Lukyanchenko[2], I. Machulin[2,21], G. Manuzio[9], S. Marcocci[7,14,31], J. Martyn[23], E. Meroni[4], M. Meyer[24], L. Miramonti[4], M. Misiaszek[19], V. Muratova[13], B. Neumair[1], L. Oberauer[1], B. Opitz[6], V. Orekhov[2], F. Ortica[25], M. Pallavicini[9], L. Papp[1], Ö. Penek[3,22], N. Pilipenko[13], A. Pocar[26], A. Porcelli[23], G. Raikov[2], G. Ranucci[4], A. Razeto[7], A. Re[4], M. Redchuk[3,22], A. Romani[25], R. Roncin[7,16], N. Rossi[7,32], S. Schönert[1], D. Semenov[13], M. Skorokhvatov[2,21], O. Smirnov[15], A. Sotnikov[15], L. F. F. Stokes[7], Y. Suvorov[2,27,33], R. Tartaglia[7], G. Testera[9], J. Thurn[24], M. Toropova[2], E. Unzhakov[13], F. L. Villante[7,28], A. Vishneva[15], R. B. Vogelaar[10], F. von Feilitzsch[1], H. Wang[27], S. Weinz[23], M. Wojcik[19], M. Wurm[23], Z. Yokley[10], O. Zaimidoroga[15], S. Zavatarelli[9], K. Zuber[24] & G. Zuzel[19]

[1]Physik-Department and Excellence Cluster Universe, Technische Universität München, Garching, Germany. [2]National Research Centre Kurchatov Institute, Moscow, Russia. [3]Institut für Kernphysik, Forschungszentrum Jülich, Jülich, Germany. [4]Dipartimento di Fisica, Università degli Studi e INFN, Milano, Italy. [5]Chemical Engineering Department, Princeton University, Princeton, NJ, USA. [6]Institut für Experimentalphysik, Universität Hamburg, Hamburg, Germany. [7]INFN, Laboratori Nazionali del Gran Sasso, Assergi, Italy. [8]Physics Department, Princeton University, Princeton, NJ, USA. [9]Dipartimento di Fisica, Università degli Studi e INFN, Genova, Italy. [10]Physics Department, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. [11]Lomonosov Moscow State University Skobeltsyn Institute of Nuclear Physics, Moscow, Russia. [12]Department of Physics and Astronomy, University of Hawaii, Honolulu, HI, USA. [13]St Petersburg Nuclear Physics Institute, NRC Kurchatov Institute, Gatchina, Russia. [14]Gran Sasso Science Institute, L'Aquila, Italy. [15]Joint Institute for Nuclear Research, Dubna, Russia. [16]AstroParticule et Cosmologie, Univ. Paris Diderot, CNRS/IN2P3, CEA/IRFU, Observatoire de Paris, Sorbonne Paris Cité, Paris, France. [17]Department of Physics, University of Houston, Houston, TX, USA. [18]Laboratorio Subterráneo de Canfranc, Canfranc Estacion Huesca, Spain. [19]M. Smoluchowski Institute of Physics, Jagiellonian University, Krakow, Poland. [20]Kiev Institute for Nuclear Research, Kiev, Ukraine. [21]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russia. [22]RWTH Aachen University, Aachen, Germany. [23]Institute of Physics and Excellence Cluster PRISMA, Johannes Gutenberg Universität Mainz, Mainz, Germany. [24]Department of Physics, Technische Universität Dresden, Dresden, Germany. [25]Dipartimento di Chimica, Biologia e Biotecnologie, Università degli Studi e INFN, Perugia, Italy. [26]Amherst Center for Fundamental Interactions and Physics Department, University of Massachusetts, Amherst, MA, USA. [27]Physics and Astronomy Department, University of California Los Angeles (UCLA), Los Angeles, California, USA. [28]Dipartimento di Scienze Fisiche e Chimiche, Università dell'Aquila, L'Aquila, Italy. [29]Present address: Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid, Spain. [30]Present address: Physics Department, University of California, San Diego, CA, USA. [31]Present address: Fermi National Accelerator Laboratory (FNAL), Batavia, IL, USA. [32]Present address: Dipartimento di Fisica, Sapienza Università di Roma e INFN, Rome, Italy. [33]Present address: Dipartimento di Fisica, Università degli Studi Federico II e INFN, Naples, Italy. *e-mail: spokesperson-borex@lngs.infn.it

METHODS

The Borexino detector. Borexino is a large liquid-scintillator experiment located deep underground at the Laboratori Nazionali del Gran Sasso in Italy. Borexino is designed to achieve extremely low background conditions. The active core of the detector consists of about 300 t of pseudocumene (1,2,4-trimethylbenzene) doped with 1.5 g per litre of PPO (2,5-diphenyloxazole) and contained in a spherical nylon inner vessel (radius $R = 4.25$ m). The scintillator is surrounded by a non-scintillating pseudocumene-based buffer liquid which serves as a shield against external radioactivity (see Extended Data Fig. 1). The scintillator fluorescence light is collected by 2,212 photomultiplier tubes mounted on the Stainless Steel Sphere (radius $R = 6.9$ m). The entire detector is enclosed in a domed, cylindrical tank filled with high-purity water, equipped with 208 photomultiplier tubes, which provides extra shielding against external radioactivity (photons and neutrons), and also serves as an active water Cherenkov veto against residual cosmic muons. A detailed description of the Borexino detector is found in ref. [26].

The SSM and the solar metallicity controversy. The SSM is a solution of the stellar evolution equations for stars of one solar mass, calibrated to match present-day, measured surface properties of the Sun. A fundamental assumption is that the Sun was initially chemically homogeneous and that during its 4.56-Gyr-long evolution, it has modified its chemical composition solely due to nuclear reactions and elemental diffusion. The model calibration is done by adjusting the mixing length parameter and the initial chemical composition in order to reproduce the observed solar luminosity, radius, and current surface composition. As a result of this procedure, the SSM has no free parameters and completely determines the mechanical and thermal properties of the Sun.

The SSM predicts that most of the solar energy (>99%) is produced by the so-called *pp* chain (see Fig. 1) that fuses hydrogen into ^{4}He: the chain is initiated by the *pp* fusion reaction and, to a minor extent, by the alternative three-body *pep* process. These reactions produce deuterons, which are efficiently converted into ^{3}He by the subsequent deuteron–proton reaction. The *pp* chain mostly terminates with the $^3\mathrm{He} + {}^3\mathrm{He} \to {}^4\mathrm{He} + 2p$ reaction (*pp*-I termination). In the late 1950s, the cross-section for the competing $^3\mathrm{He} + {}^4\mathrm{He} \to {}^7\mathrm{Be} + \gamma$ reaction was discovered[40] to be about one thousand times larger than previously thought, causing the branching ratios of the *pp*-II and *pp*-III terminations to be non-negligible. An alternative process is the so-called CNO cycle, a closed-loop nuclear reaction in which ^{12}C, ^{14}N, and ^{16}O nuclei catalyse hydrogen fusion into ^{4}He. The CNO cycle is a subdominant energy-producing mechanism in stars like the Sun or lighter, but is believed to be the dominant fusion mechanism in heavier or older stars.

For each ^{4}He nucleus produced in the Sun, two electron-flavour neutrinos are emitted. Neutrinos free-stream across the solar plasma and reach the Earth travelling close to the speed of light in about 8 min, resulting in a total flux of about $6.5 \times 10^{10}\,\mathrm{cm^{-2}\,s^{-1}}$. The solar-neutrino spectrum depends on the branching ratios of the different *pp* chain terminations and on the relative intensity of the *pp* chain and the CNO cycle. A large percentage (about 90%) of the neutrinos emitted by the Sun are produced in the primary *pp* fusion reaction (producing *pp* neutrinos). Most of the remaining 10% of the solar-neutrino flux is emitted in the electron capture reaction on ^{7}Be (producing ^{7}Be neutrinos), which appears along the *pp*-II branch of the chain. Smaller contributions come from *pep* fusion (the *pep* neutrinos) and from ^{8}B decays in the *pp*-III branch (producing ^{8}B neutrinos). Neutrinos from proton capture of ^{3}He (hep neutrinos) are expected to be emitted with negligible probability (10^{-7}) and are beyond current detection sensitivity. The predicted energy spectrum of all neutrinos emitted along the *pp* chain, including spectral shapes and intensity before neutrino oscillations are shown in Fig. 1.

The predictions of the SSM have been tested by solar-neutrino experiments and by helioseismology (which determines the properties of the solar interior by studying the propagation of seismic waves at the Sun's surface). However, important questions about the Sun still call for an answer. For example, the solar metallicity—the abundance of elements heavier than He—is poorly understood, although it is a fundamental input when constructing SSMs and a relevant parameter in astrophysics, since almost all determinations of elemental abundances in astronomical objects rely upon the solar composition. Recent determinations of the solar surface composition[41–43] suggest that the solar metallicity might be lower than previously assumed[44,45]. SSMs that incorporate these lower abundances, however, agree less well with helioseismic data: this is often referred to as the solar metallicity problem.

Solar-neutrino measurements provide fundamental clues for the solution of this puzzle. Indeed, the opacity of the solar plasma is strongly influenced by the presence of heavy elements. Since opacity determines the efficiency of radiative energy transfer, the metal content of solar matter affects the temperature profile of the Sun. As a consequence, metallicity determines the branching ratios for the various terminations of the *pp* chain, as well as the relative intensity of the *pp* chain with respect to the CNO cycle. A precise determination of the solar-neutrino fluxes comprising both the *pp* chain and the CNO cycle is thus a direct, robust way to settle the solar metallicity controversy. In the main text we compare our experimental results with predictions of HZ and LZ SSMs[18].

Neutrino oscillations and the MSW effect. For many years, the experimental results on solar neutrinos have been at odds with the predictions of the SSM: all the experiments observed a large deficit of neutrinos with respect to expectations. This 30-year-long controversy was settled only in 2002 by the experiment SNO[11], which proved unambiguously that the solution to the 'solar-neutrino problem' was not to be searched for in solar physics, but in neutrino physics, namely, in the quantum mechanics phenomenon of flavour oscillations[12]. Through this mechanism, solar neutrinos, which are born in the Sun as electron neutrinos, ν_e, have a non-zero probability to transform into neutrinos with a different flavour (either ν_μ or ν_τ) during propagation and are therefore less likely to be detected on Earth. For oscillations to occur, two conditions must be met: (1) mass and flavour eigenstates for neutrinos must not coincide, which implies the existence of a non-trivial mixing matrix which transforms one into the other; and (2) the mass of at least one neutrino must be different from 0. The relevant parameters for solar-neutrino oscillations are the mixing angle θ_{12} and the squared mass difference between the mass eigenstates, mostly contributing to ν_e, that is, Δm^2_{12}. The probability of flavour conversion is enhanced when neutrinos cross the dense solar medium, because of coherent forward scattering on electrons. This mechanism is referred to as the Mikheyey–Smirnov–Wolfenstein (MSW) matter effect[13,14] and for the specific values of Δm^2_{12} and of the Sun density profile it fully describes solar neutrinos with energies greater than about 5 MeV. For energies below 1 MeV, the vacuum oscillation mechanism dominates, whereas in the intermediate-energy region, a smooth transition occurs. Figure 3 shows the survival probability P_{ee} for ν_e produced in the Sun as a function of the neutrino energy (pink curve) for the oscillation parameters obtained by a global fit to all solar-neutrino, reactor and accelerator experiments[19]. The values of Δm^2_{12} (about $7.5 \times 10^{-5}\,\mathrm{eV}^2$) and of θ_{12} (about 33°) correspond to the so-called Large Mixing Angle solution (LMA) of the solar-neutrino problem.

Event selection and residual backgrounds. The analysis starts with data selection aimed at reducing the rate of background events. The selection criteria, conceptually similar for the LER and HER, are conceived to: (1) reject cosmic muons penetrating the mountain shield; (2) reduce the cosmogenic background, that is, the decays of short-lived radioactive elements produced in muon-induced nuclear spallation processes in the detector; and (3) select a fiducial volume of the scintillator, optimized separately for the LER and HER-I/II analyses.

Rejection of muons is achieved by combining the external Cherenkov veto information with a pulse shape analysis of the scintillator signals, and displays an overall efficiency[30] of 99.992%. The reduction of cosmogenic background is obtained by excluding events collected during a given time Δt following every muon crossing the scintillator.

For the LER, a short muon veto time $\Delta t = 300$ ms is enough to efficiently suppress most relevant cosmogenic isotopes. An exception is ^{11}C ($Q = 0.96$ MeV, β^+, $\tau = 29.4$ min), which is produced in situ by muon spallation, and has a mean lifetime that greatly exceeds the short muon veto time cut. ^{11}C has a fairly constant concentration in the scintillator (around 30 counts per day per 100 t) determined by the equilibrium between its production and decay rate and cannot be reduced by any purification procedure. It is therefore one of the most important backgrounds and must be treated with a specific analysis (see next paragraph).

For the HER, the rejection of cosmogenic background requires a larger time window of $\Delta t = 6.5$ s to suppress ^{12}B, ^{8}He, ^{9}C, ^{9}Li, ^{8}B, ^{6}He and ^{8}Li decays. Furthermore, for the HER analysis a 2-ms veto is applied after muons that cross the buffer liquid only. This veto aims at rejecting 4.95-MeV γ-rays following the capture of cosmogenic neutrons on ^{12}C nuclei; an additional cut is applied around the capture position of cosmogenic neutrons, when this happens inside the scintillator, to remove ^{10}C ($Q = 3.6$ MeV, β^+, $\tau = 27.8$ s).

Both in the LER and in the HER, ^{214}Bi and ^{214}Po from the ^{238}U natural decay chain are removed by exploiting the space-time correlation of their fast $\beta + \alpha$ delayed coincidence decays.

The analysis in the LER and HER-I/II use different fiducial volumes. The LER fiducial volume focuses on suppressing external γ-rays from ^{40}K, ^{214}Bi and ^{208}Tl contained in materials surrounding the scintillator. It consists of the central 71.3 t of scintillator, selected by applying a radial cut ($R < 2.8$ m) and a cut along the vertical axis ($-1.8\,\mathrm{m} < z < 2.2$ m). The HER is above the energy of the aforementioned γ-rays. The analysis in HER-I only requires a $z < 2.5$ m cut to suppress background events related to a small pinhole in the nylon vessel that causes scintillating fluid to leak into the region surrounding it. The total HER-I target mass is 227.8 t. The analysis in HER-II uses the entire scintillator volume of 266 t. More details on the selection criteria can be found in refs [24,25,27].

After the selection cuts described above, some residual background remains both in the LER and in the HER. The LER residual background is detailed in Table 1, and is mostly due to traces of radioactive isotopes contaminating the scintillator—that is, ^{14}C, ^{210}Po (either from ^{210}Pb decay or out of equilibrium), ^{85}Kr, ^{210}Bi (from ^{210}Pb) and pile-up of uncorrelated events. A small contribution to the LER rate also comes from external ^{208}Tl, ^{214}Bi and ^{40}K γ-rays emerging from materials surrounding the scintillator. In the LER fit, the ^{14}C rate is

quantified and constrained using an independent sample of events acquired without any trigger threshold[25]. The contribution of pile-up, dominated by simultaneous ^{14}C decays at different detector positions, is treated using the following two methods described in refs [25,29]: in one case, we construct the pile-up spectrum starting from real or Monte Carlo datasets; in the other, we convolve all spectral components with a randomly acquired spectrum (that is, with events acquired with a solicited, external trigger).

The residual backgrounds affecting the HER-I and HER-II are also listed in Table 1. Some of the internal events (that is, events uniformly distributed in the scintillator volume) are due to muons, cosmogenic isotopes, and ^{214}Bi decays surviving the cuts. The total contribution of these backgrounds has been evaluated separately for the HER-I and the HER-II, following the procedure described in ref. [24], and constrained in the fit. In addition, the presence of untagged ^{11}Be ($Q = 11.5$ MeV, β^-, $\tau = 19.9$ s) is estimated by adopting a technique based on a multivariate fit, which includes the energy spectrum and the time profile of events with respect to the preceding muon, and is found to be compatible with zero. The HER-I is also affected by internal ^{208}Tl decays, which come from the residual ^{232}Th contamination of the liquid scintillator. In the fit, this rate is constrained to the value obtained by counting the ^{212}Bi–^{212}Po $\beta + \alpha$ fast delayed coincidences. External ^{208}Tl contamination contributes to the HER-I with two distinct components: one from contamination directly on the inner vessel surface, and another from decays of nuclei that have recoiled off the inner vessel into the liquid scintillator or originated from the volatile progenitor of ^{208}Tl, ^{220}Rn, which has emanated out of the nylon. The rates of both components are left free to vary in the radial fit. Finally, HER-I and HER-II are also polluted by γ-rays following the capture of radiogenic neutrons produced via (α, n) or spontaneous fission reactions of ^{238}U, ^{235}U and ^{232}Th in the Stainless Steel Sphere and photomultiplier tubes. This rate is also a free parameter of the fit.

The ^{11}C background. The ^{11}C background is not removed by the short veto cut after muons. To disentangle its contribution from the neutrino signal, we use the TFC method[23,27], which exploits the time and space correlation between muons, the neutrons they produce in combination with ^{11}C, and the subsequent ^{11}C decays. With this method we divide the events passing the selection cuts in two complementary datasets: one is depleted in ^{11}C (TFC-subtracted) and preserves (64.28 ± 0.01)% of the total exposure; the other contains (92 ± 4)% of the ^{11}C (TFC-tagged). The energy spectra of these two datasets are fitted simultaneously in the multivariate fit (see next paragraph). The residual ^{11}C (positron) background in the TFC-subtracted spectrum is further disentangled from electron-like events by including in the multivariate fit the distribution of a pulse-shape discrimination variable[23,27]. It is in fact observed that the time distribution of scintillation photons slightly differs between electron and positron events, for the following reasons: (1) positron produces ortho-positronium half of the time, which delays the annihilation by around 3 ns (ref. [46]); (2) the positron energy deposition occurs in multiple sites within the detector, owing to the production of annihilation γ-rays. These effects tend to delay and extend the time distribution of the scintillator pulse for positrons with respect to electron events, a handle we exploit for ^{11}C background rejection.

Fitting procedure for extraction of solar-neutrino rates. To disentangle the neutrino signal rates from the residual background, we apply different fitting strategies for the LER and the HER. For LER, we adopt a multivariate approach and simultaneously fit the TFC-subtracted and the TFC-tagged energy spectra, the spatial distribution, and the distribution of the pulse-shape discrimination variable. The spatial distribution is crucial to separate the residual external background component, while the pulse-shape estimator is optimized to separate positrons from electrons, which is key to disentangling ^{11}C from the other fit species (see above). The reference radial distributions for external and internal events used in the multivariate fit are built with a comprehensive Geant4-based Monte Carlo simulation, carefully tuned and validated with calibration data[28,29]. The spectral shapes of signal and background components used in the multivariate fit of the LER are also obtained from simulations. In addition, the fit of the energy spectra is performed using analytical spectral functions[25,27], where the nonlinearity of the energy scale (due, for example, to ionization quenching and Cherenkov light emission) and the spatially non-uniform detector response are included via nuisance parameters, some of which are left free to vary in the fit. The reference positron pulse-shape distribution used in the LER multivariate fit is based on events selected with the TFC method described above, tuned to obtain a nearly pure sample of ^{11}C events. The reference electron pulse-shape distribution is obtained from simulations and checked on data using electron-like events isolated via the ^{214}Bi–^{214}Po coincidences.

In the HER-I and HER-II, the analysis is based on a fit to the radial distribution of the events to separate the ^{8}B neutrino signal (uniformly distributed in the scintillator) from the external background components. Like the LER fit, the reference radial distributions for external and internal events used in the HER fit are built with Geant4-based Monte Carlo simulations. For more details on the fit to extract the neutrino signal see ref. [27].

Systematic uncertainties in the analysis. The detector energy response and uniformity has been carefully studied by means of an extensive calibration campaign which was carried out in 2009[28]. The calibration data were used to tune the input parameters of the Borexino Monte Carlo package, a custom Geant4-based code[47] that can simulate all processes following the interaction of a particle in the detector, including all known characteristics of the apparatus[29]. After tuning, the agreement between Monte Carlo and calibration data is very good for both the LER and the HER: for the energies relevant to the LER analysis, the overall uncertainty is below 1%, while for the HER analysis, it is around 1.9%.

In spite of this remarkable understanding of the detector response throughout the scintillator volume and in a large energy range, an extensive study of possible sources of systematic errors has been performed both for the LER and for the HER. The results of these studies are summarized in Extended Data Tables 1 and 2, respectively.

Concerning the analysis in the LER, the main contribution to the systematic error comes from the fit model, that is, possible residual inaccuracies in the modelling of the detector response (energy scale, uniformity of the energy response, pulse-shape discrimination shape) and uncertainties in the theoretical energy spectra used in the fit. These systematic effects have been estimated by means of a Monte Carlo method: an ensemble of 100,000 datasets are simulated from a family of probability density functions, which includes deformations caused by the inaccuracies under study. The magnitude of the deformations was chosen to be within the range allowed by the available calibration data. These data are then fitted following the same procedure used for real data and differences in the results are quoted as systematics (see first line in Extended Data Table 1).

The second source of systematics is related to the fit method, that is, whether the reference probability density functions used in the fit are entirely derived from Monte Carlo simulations or analytically. Further systematic effects arise from the choice of the energy estimator, from the details of the implementation of the pile-up of uncorrelated events, from using different fit energy ranges and binning, from the inclusion of an independent constraint on ^{85}Kr obtained from its sub-dominant delayed coincidence decay (branching ratio 0.43%), and from the estimation of the target fiducial mass. This last uncertainty is determined with calibration data, by using sources deployed in known positons throughout the detector volume.

Concerning the HER analysis, the most important systematic uncertainties arise from the determination of the target mass, from the energy scale, and from the z-cut applied in the HER-I range (see Extended Data Table 2).

The target mass uncertainty is related to the fact that the amount of scintillator contained in the inner vessel is slowly decreasing (by less than 0.5 m^3 per year), due to a small pinhole in the nylon membrane. We monitor the evolution of the scintillator mass on a week-by-week basis, by studying the inner vessel shape, which is obtained from the spatial distribution of its surface contamination. This method gives an average total mass of 266 t with an error of about 2%.

The impact of the uncertainty of the energy scale on the number of events falling in the HER-I and HER-II energy window has been evaluated with a full Monte Carlo simulation and has been included in the systematic error (see second line of Extended Data Table 2).

As mentioned in the main text, the HER-I analysis requires a cut on the vertical coordinate to remove background events owing to a small pinhole in the nylon vessel that causes the scintillator to leak into the buffer liquid. To estimate possible systematics associated to this cut, the HER-I analysis was performed with a modified z-cut, ±0.5 m around the chosen value (2.5 m). Differences in the results have been included as systematic error.

Frequentist hypothesis test of MSW versus vacuum oscillations. Borexino provides results on the electron neutrino survival probability (P_{ee}) in the entire solar-neutrino energy range. We are therefore able to perform a statistical study to compare the compatibility of our measurement with two different hypotheses: the standard oscillation scenario, MSW-LMA, and the vacuum-LMA scenario, in which matter effects are not present (and which is taken as our null hypothesis).

The survival probability $P_{ee}^{\text{MSW-LMA}}$ in the MSW-LMA scenario depends not only on the oscillation parameters θ_{12}, θ_{13} and Δm^2_{12} valid in vacuum, but also on the neutrino-energy-dependent potential characterizing the interaction of neutrinos with the dense solar core. It can be expressed as follows[48]:

$$P_{ee}^{\text{MSW-LMA}} = \frac{1}{2}\cos^4\theta_{13}\left(1 + \cos2\theta_{12}^{\text{M}}\cos2\theta_{12}\right)$$

where

$$\cos2\theta_{12}^{\text{M}} = \frac{\cos2\theta_{12} - \beta}{\sqrt{(\cos2\theta_{12} - \beta)^2 + \sin^2 2\theta_{12}}}$$

and

$$\beta = \frac{2\sqrt{2}\,G_{\text{F}}\cos^2\theta_{13} n_e E_\nu}{\Delta m_{12}}$$

where θ^{M} is the mixing angle in matter, G_F is the Fermi coupling constant and n_e is the density of electrons in the matter. Using the current set of oscillation parameters and errors derived in ref. [19], and following the procedure described in ref. [27], we obtain the pink band in Fig. 3.

If matter effects were not present, the survival probability for solar neutrinos would be approximated by the expression P_{ee}^{vacuum}:

$$P_{ee}^{\text{vacuum}} = \cos^4\theta_{13}\left(1-\frac{1}{2}\sin^2 2\theta_{12}\right)+\sin^4\theta_{13}$$

which is independent of the neutrino energy E_ν. Taking for θ_{13} and θ_{12} the values and errors measured by reactor neutrino experiments in refs [38,39], the survival probability P_{ee}^{vacuum} as a function of E_ν corresponds to the grey band in Fig. 3.

We performed a frequentist analysis, in which we adopt a test statistics t based on the ratio between the likelihood L obtained assuming MSW-LMA and vacuum-LMA:

$$t = -2\log[L(\text{MSW})/L(\text{vacuum})] = \chi^2(\text{MSW}) - \chi^2(\text{vacuum})$$

The probability distribution of t is built with a Monte Carlo method: we randomly generate thousands of values of P_{ee} in the MSW-LMA hypothesis (by sampling the pink curve in Fig. 3 and including both theoretical and experimental uncertainties) and for each set of data we estimate t and build its distribution (red curve on the left in Extended Data Fig. 2). In the same way, we simulate thousands of P_{ee} values in the vacuum-LMA hypothesis and we build the corresponding t distribution (blue curve on the right in Extended Data Fig. 2).

The actual Borexino results for P_{ee} for *pp*, ^{7}Be, *pep* and ^{8}B gives a value of $t_{BX} = -4.16$ (indicated as a dashed line in Extended Data Fig. 2), which allows us to disfavour the vacuum-LMA hypothesis with a P value of 0.018 (integral of the small tail of the blue curve to the left of t_{BX}), corresponding to a C.L. of 98.2%. For more details on the choice of the test statistics see ref. [49].

Frequentist and Bayesian hypothesis test of the HZ versus LZ models. The combination of the Borexino measurement on ^{8}B and ^{7}Be fluxes provides an interesting hint in favour of the solar temperature profile predicted by the HZ-SSM. This was obtained by performing both a frequentist and a Bayesian hypothesis test.

In the frequentist analysis, we used a test statistics t based on the ratio between the likelihood obtained assuming HZ and LZ:

$$t = -2\log[L(\text{HZ})/L(\text{LZ})] = \chi^2(\text{HZ}) - \chi^2(\text{LZ})$$

The probability distribution of t is built with a Monte Carlo method (full Neumann construction of the confidence intervals): we randomly generate thousands of fake ^{7}Be–^{8}B results in the HZ hypothesis (sampling a distribution that includes both theoretical and experimental errors) and for each set of data we estimate t (red distribution on the left in Extended Data Fig. 3). In the same way, we simulate thousands of fake ^{7}Be–^{8}B results in the LZ hypothesis and we build the corresponding t distribution (in blue on the right in Extended Data Fig. 3).

The value of t corresponding to the actual Borexino result for ^{7}Be–^{8}B is shown in the plot as the dotted line at $t_{BX} = -3.49$, relatively far from the maximum of the LZ probability distribution (blue curve). This allows us to disfavour the LZ hypothesis with a P value of 0.034 (integral of the small tail of the blue curve to the left of t_{BX}), corresponding to a C.L. of 96.6%. The result is slightly better than the median P value expected (0.058), which corresponds to a median significance of 94.2% C.L. For more details on the choice of the test statistics see ref. [49].

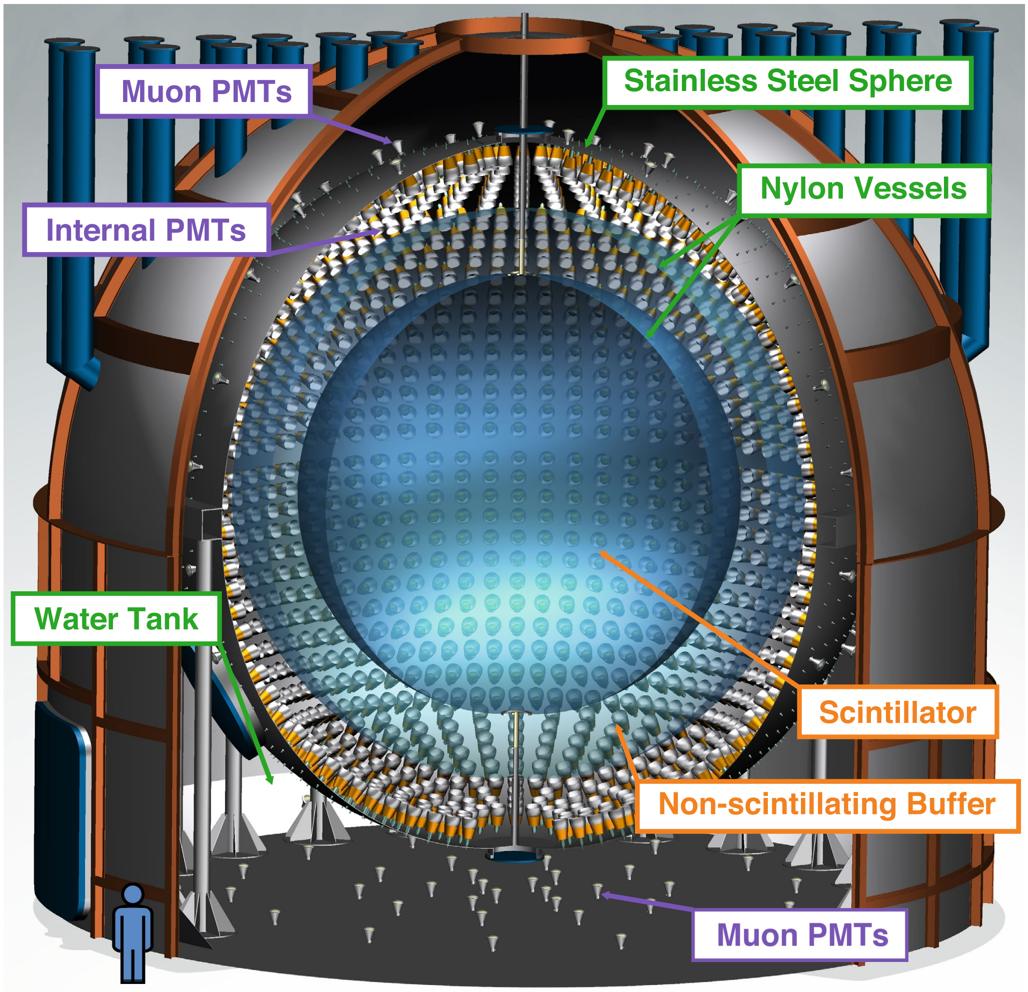
In the Bayesian analysis we constructed two models, one for the HZ and the other for the LZ hypothesis, in which the free parameters are the fluxes of ^{8}B and of ^{7}Be. The model predictions are used as prior probability distributions. The likelihood is constructed as the sum of two Gaussian measurements, one for the flux of ^{8}B and the other for the flux of ^{7}Be.

We compare the two models assuming that they have the same probability a priori (50% for the HZ hypothesis and 50% for LZ hypothesis). Like the frequentist analysis, the data show a mild preference for HZ with respect to LZ. The odds are 5:1 or, equivalently, the Bayes factor is 4.9. For more details on the Bayesian method see ref. [37].

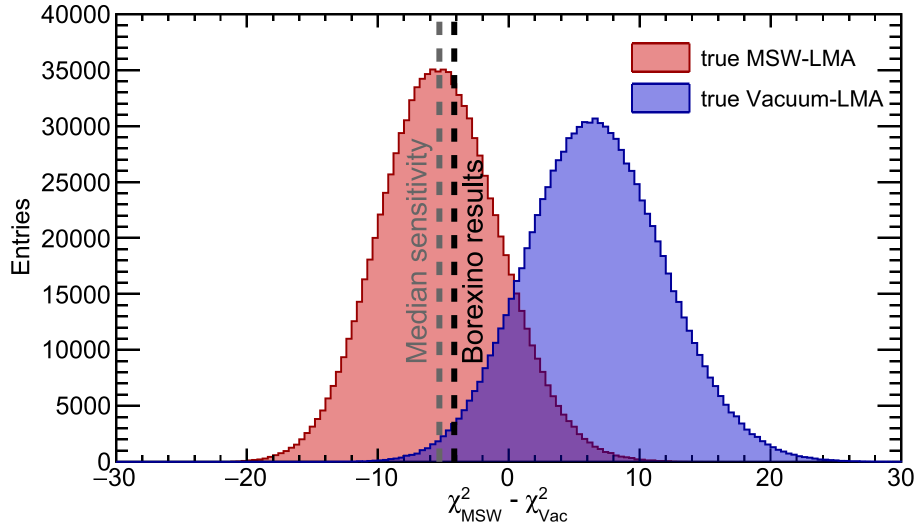
Data availability

The datasets generated during the current study are freely available in the repository https://bxopen.lngs.infn.it/. Additional information is available from the Borexino Collaboration spokesperson (spokesperson-borex@lngs.infn.it) upon reasonable request.

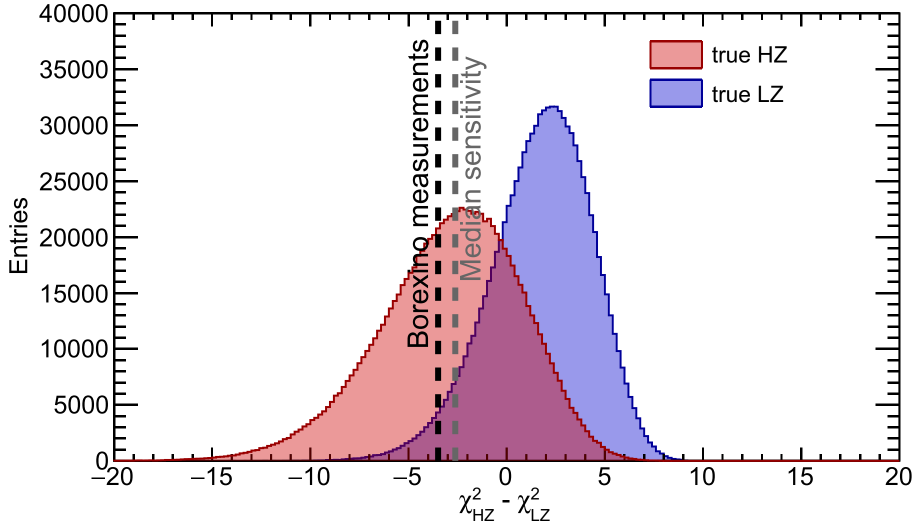
40. Holmgren, H. & Johnston, R. $He^3(\alpha,\gamma)Li^7$ and $He^3(\alpha,\gamma)Be^7$ reactions. *Phys. Rev.* **113**, 1556 (1959).
41. Asplund, M., Grevesse, N., Sauval, A. J. & Scott, P. The chemical composition of the Sun. *Annu. Rev. Astron. Astrophys*. **47**, 481 (2009).
42. Caffau, E., Ludwig, H. G., Steffen, M., Freytag, B. & Bonifacio, P. Solar chemical abundances determined with a CO5BOLD 3D model atmosphere. *Sol. Phys*. **268**, 255 (2011).
43. Asplund, M., Grevesse, N. & Sauval, A. J. *The Solar Chemical Composition*. (eds Barnes, T. G. & Bash, F. N.) Astronomical Society of the Pacific Conference Series 336, 25 (ASP, 2005).
44. Grevesse, N. & Sauval, A. J. Standard solar compositon. *Space Sci. Rev*. **85**, 161 (1998).
45. Grevesse, N. & Noels, A. in *Origin and Evolution of the Elements* (eds Prantzos, N., Vangioni-Flam, E. & Casse, M.) 15 (Cambrige Univ. Press, Cambridge, 1993).
46. Franco, D., Consolati, G. & Trezzi, D. Positronium signature in organic liquid scintillators for neutrino experiments. *Phys. Rev. C* **83**, 015504 (2011).
47. Geant4. A simulation toolkit. https://geant4.web.cern.ch/ (2018).
48. Bahcall, J. N. & Pena-Garay, C. Solar models and solar neutrino oscillations. *New J. Phys*. **6**, 63 (2004).
49. Blennow, M. & Coloma, P. Quantifying the sensitivity of oscillation experiments to the neutrino mass ordering. *J. High Energy Phys*. **03**, 028 (2013).



Extended Data Fig. 1 | The Borexino detector. Schematic view of the 'onion-like' structure of the Borexino apparatus. From outside to inside: the external water tank; the Stainless Steel Sphere, where about 2,200 photomultiplier tubes (PMTs) are mounted; the outermost nylon vessel, which serves as a barrier against radon; the innermost nylon vessel, which contains 300 t of liquid scintillator, the active detection medium.



Extended Data Fig. 2 | Frequentist hypothesis test of MSW-LMA versus vacuum-LMA. The probability distribution of the test statistics *t* is obtained by simulating thousands of sets of P_{ee} values (at the *pp*, ^{7}Be, *pep* and ^{8}B energies) in the MSW-LMA hypothesis (red curve on the left) and in the vacuum-LMA hypothesis (blue curve on the right). The dotted black line corresponds to the results of Borexino discussed in the main text.



Extended Data Fig. 3 | Frequentist hypothesis test for LZ and HZ. The probability distribution of the test statistics *t* is obtained by simulating thousands of fake sets of ^{8}B–^{7}Be values in the HZ hypothesis (red curve on the left) and in the LZ hypothesis (blue curve on the right). The dotted black line corresponds to the results of Borexino discussed in the main text.

Extended Data Table 1 | LER analysis systematics

Source of uncertainty	*pp* neutrinos		^{7}Be neutrinos		*pep* neutrinos	
	-%	+%	-%	+%	-%	+%
Fit models (see text)	-4.5	+0.5	-1.0	+0.2	-6.8	+2.8
Fit method (analytical/Monte Carlo)	-1.2	+1.2	-0.2	+0.2	-4.0	+4.0
Choice of the energy estimator	-2.5	+2.5	-0.1	+0.1	-2.4	+2.4
Pile-up modeling	-2.5	+0.5	0	0	0	0
Fit range and binning	-3.0	+3.0	-0.1	+0.1	-1.0	+1.0
Inclusion of the ^{85}Kr constraint	-2.2	+2.2	0	+0.4	-3.2	0
Live time	-0.05	+0.05	-0.05	+0.05	-0.05	+0.05
Scintillator density	-0.05	+0.05	-0.05	+0.05	-0.05	+0.05
Fiducial volume	-1.1	+0.6	-1.1	+0.6	-1.1	+0.6
Total systematics (%)	**-7.1**	**+4.7**	**-1.5**	**+0.8**	**-9.0**	**+5.6**

Relevant sources of systematic uncertainties and their contributions to the measured neutrino interaction rates for the LER analysis.

Extended Data Table 2 | HER analysis systematics

Source of uncertainty	*HER-I* -%	*HER-I* +%	*HER-II* -%	*HER-II* +%	*HER* (tot) -%	*HER* (tot) +%
Target mass	-2.0	+2.0	-2.0	+2.0	-2.0	+2.0
Energy scale	-0.5	+0.5	-4.9	+4.9	-1.7	+1.7
z-cut	-0.7	+0.7	0	0	-0.4	+0.4
Live time	-0.05	+0.05	-0.05	+0.05	-0.05	+0.05
Scintillator density	-0.05	+0.05	-0.05	+0.05	-0.05	+0.05
Total systematics (%)	**-2.2**	**+2.2**	**-5.3**	**+5.3**	**-2.7**	**+2.7**

Relevant sources of systematic uncertainties and their contributions to the measured neutrino interaction rates for the HER analyses.

https://doi.org/10.1038/s41586-018-0607-z

Rock fluidization during peak-ring formation of large impact structures

Ulrich Riller[1]*, Michael H. Poelchau[2], Auriol S. P. Rae[3], Felix M. Schulte[1], Gareth S. Collins[3], H. Jay Melosh[4], Richard A. F. Grieve[5], Joanna V. Morgan[3], Sean P. S. Gulick[6,7], Johanna Lofi[8], Abdoulaye Diaw[8], Naoma McCall[6,7], David A. Kring[9] & IODP–ICDP Expedition 364 Science Party[10]

Large meteorite impact structures on the terrestrial bodies of the Solar System contain pronounced topographic rings, which emerged from uplifted target (crustal) rocks within minutes of impact. To flow rapidly over large distances, these target rocks must have weakened drastically, but they subsequently regained sufficient strength to build and sustain topographic rings. The mechanisms of rock deformation that accomplish such extreme change in mechanical behaviour during cratering are largely unknown and have been debated for decades. Recent drilling of the approximately 200-km-diameter Chicxulub impact structure in Mexico has produced a record of brittle and viscous deformation within its peak-ring rocks. Here we show how catastrophic rock weakening upon impact is followed by an increase in rock strength that culminated in the formation of the peak ring during cratering. The observations point to quasi-continuous rock flow and hence acoustic fluidization as the dominant physical process controlling initial cratering, followed by increasingly localized faulting.

Large hypervelocity impact structures show a distinct size–morphology progression[1] (Fig. 1), which depends on the gravity and target rock type of the impacted body. In this regard, the study of internal topographic rings—so-called peak rings[2]—are of particular importance in understanding the formation of peak-ring impact structures (Fig. 1b) and multi-ring impact basins (Fig. 1c)[3]. As crater diameter increases beyond the maximum size of a bowl-shaped crater, the depth-to-diameter ratio of the crater decreases. On Earth, peak-ring crater formation (Fig. 2, Supplementary Information) takes place in minutes[1,4] and implies extreme deformation rates accompanying large displacements. Peak-ring craters can be a few hundred kilometres in diameter, yet merely a few kilometres deep, with the peak rings greatly elevated above crater floors. To explain this topographic characteristic, peak-ring crater formation requires drastic mechanical weakening of the target rocks. Weakening is thought to be caused by a decrease in the angle of internal friction and cohesion and results in large-scale fluid-like behaviour of target rock during part of the cratering process[4–7]. Towards the end of the cratering process, however, rock strength needs to be sufficiently high to form and sustain topographically elevated peak rings.

A number of mechanisms for target-rock weakening have been proposed. These include impact-induced fracturing and fragmentation of the target rocks[8–16], wholesale thermal softening by shock heating[6], fault weakening[17] by shear heating[18] or other processes, and acoustic fluidization[19,20]. In this last process, short-wavelength, high-frequency pressure oscillations around the lithostatic pressure temporarily reduce the overburden pressure and, thus, friction between fractured target rocks.

Because direct observations are extremely limited, the exact mechanisms and duration of target rock weakening during large impact cratering are unknown. In particular, unequivocal physical evidence for acoustic fluidization or fault weakening in large impact craters remains to be identified. Large extra-terrestrial craters can only be analysed by remote sensing, which provides little or no subsurface structural information. With estimated original diameters between 200 km and 250 km, Vredefort (South Africa), Sudbury (Canada) and Chicxulub (Mexico), known as 'the big three'[21], are the largest impact structures known on Earth. Vredefort and Sudbury, however, are eroded to variable depths[22] of about 10 km and about 5 km, respectively, and so are largely missing the upper and most displaced target rocks (Fig. 2d). Chicxulub is the sole near-pristine, large impact structure with a topographic peak ring on Earth (Fig. 1d)[23–27], but post-impact sedimentary strata hundreds of metres thick have buried the impact structure, hindering direct access to the target rocks. Recent drilling by Expedition 364 of the International Ocean Discovery Program (IODP) with the International Continental Scientific Drilling Program (ICDP)[24,28] into the target rocks that constitute the peak ring at Chicxulub has now provided unprecedented insight into target rock deformation, weakening mechanisms and peak-ring formation in large-scale impact cratering.

Structural characteristics of target rock

A total of 829 m of core was recovered from Expedition 364 borehole M0077A (Fig. 1d), starting at 506 m below sea floor (m.b.s.f.)[24,28]. The recovered core includes 112 m of post-impact pelagic carbonate rock, followed by 130 m of impact melt rock and suevite, and 587 m of pervasively shocked target rock. The target rock consists of coarse-grained, alkali-feldspar-rich granitoid rock hosting uniformly oriented, pre-impact mafic and felsic sheet intrusions (Extended Data Fig. 1). At depths between 1,220 and 1,316 m.b.s.f., the target rock is mingled with impact melt rock on the decimetre to metre scale. Elsewhere in the target rock, impact melt rock is rather sparse. The mean density (2.41 g cm^{-3}) and mean P-wave velocity (4.1 km s^{-1}) of the target rock are considerably lower than those of typical felsic basement rocks (>2.6 g cm^{-3} and >5.5 km s^{-1})[24,28]. These petrophysical characteristics indicate substantial mechanical modification of the rock, notably in terms of increased porosity[29].

[1]Institut für Geologie, Universität Hamburg, Hamburg, Germany. [2]Department of Geology, Universität Freiburg, Freiburg, Germany. [3]Department of Earth Science and Engineering, Imperial College London, London, UK. [4]Department of Earth, Atmospheric and Planetary Sciences, Purdue University, West Lafayette, IN, USA. [5]Centre for Planetary Science and Exploration, Western University, London, Ontario, Canada. [6]Institute for Geophysics, University of Texas, Austin, TX, USA. [7]Department of Geological Sciences, Jackson School of Geosciences, University of Texas, Austin, TX, USA. [8]Géosciences Montpellier, CNRS, Université de Montpellier, Montpellier, France. [9]Universities Space Research Association, Lunar and Planetary Institute, Houston, TX, USA. [10]A list of participants and their affiliations appears at the end of the paper. *e-mail: ulrich.riller@uni-hamburg.de

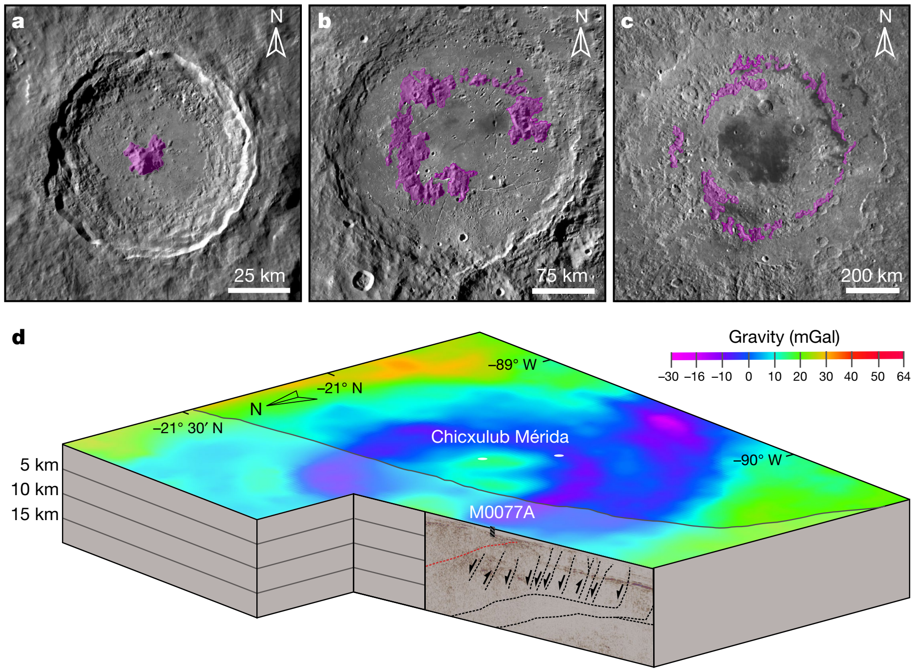


Fig. 1 | Typical impact structures on the Moon (http://quickmap.lroc.asu.edu) and the geophysical characteristics of the Chicxulub impact structure. Topographically elevated areas in **a**–**c** are highlighted in magenta. **a**, Tycho (diameter 85 km) is a central-peak crater. **b**, Schrödinger[34] (diameter 312 km) is a peak-ring impact structure. **c**, Orientale (diameter 930 km) is a multi-ring impact basin. **d**, Combined Bouguer gravity and seismic line A[27] of the Chicxulub impact structure. Offshore seismic data[27] indicate that the Chicxulub peak ring roughly correlates with a gravity low. The location of drill hole M0077A on the peak ring is indicated. Half-arrows indicate the sense of displacement on faults.

The post-impact carbonate rock is unstrained. Pre-impact deformation, however, of the granitoid target rock is evident through the sporadic presence of weak shape-preferred orientations of alkali-feldspar, plagioclase, quartz and biotite. The grain-shape alignment of these minerals formed under high-grade metamorphic conditions, as indicated by viscous deformation of feldspars and quartz[28]. Crystal-plastic strain cannot account for the reduced density and P-wave velocity of the target rock. Consequently, impact processes, including the damage caused by the passage of the shock wave, and deformation during peak-ring formation, must have caused the anomalous geophysical properties of the target rock[29].

Observed shock-induced structures in the target rock consist of shatter cones, microscopic planar deformation features and planar fractures in quartz and feldspars, as well as kinked biotite[28]. Severe structural target rock modification is evident by brittle and viscous deformation structures, including: (1) pervasive, irregular grain-scale fractures, (2) zones of cataclasite and ultra-cataclasite, (3) striated shear faults, (4) crenulated mineral foliations, and (5) brittle–ductile band structures (Figs. 3 and 4). The formation of (1) to (3) substantially increases the volume of deformed rock and thus accounts for the observed reduction in density and P-wave velocity[29].

The spatial distribution of macroscopic deformation structures indicates highly heterogeneous deformation in the target rock (Fig. 3). Microscopic inspection of the granitoid target rock reveals the pervasive presence of intra- and inter-granular dilation fractures displaying jigsaw fragment geometry (Fig. 4a). Zones of strongly comminuted material separate displaced mineral fragments (Fig. 4b). These cataclasite zones range in thickness from millimetres to centimetres (Fig. 4a–c, g) and indicate local differential shearing during cataclastic deformation. Locally, cataclasite zones grade into, or are truncated by, flow-foliated ultra-cataclasite, characterized by alternating quartz- and feldspar-rich layers (Fig. 4d). Crystal-plastic distortion of plagioclase (Fig. 4e) and quartz (Fig. 4f) indicate that the target rock accumulated some plastic strain before pervasive fracturing and cataclastic flow. Zones of (ultra-)cataclasite and crude mineral foliations, defined by the shape-preferred orientation of biotite and coarse layers of quartz and feldspars, are sporadically kinked (Fig. 4g, h). In summary, cataclastic deformation displays variable intensity throughout the cored target rock, which is evident by its localization and variable degree of comminution.

A total of 602 shear faults, with well-defined slip lineations, were recorded in the granitoid target rock (Fig. 3), with the total number of shear faults being vastly higher. By contrast, only 13 shear faults with slip lineations were identified in the post-impact carbonate rock and consist of a few millimetre-long calcite fibres (Fig. 4i), typical of seismic stick–slip faulting[30]. Slip lineations in the target rock, however, form pronounced ridges and grooves of strongly comminuted host rock material (Fig. 4j). Displacements on these faults may amount to several decimetres[28]. Although the post-impact carbonate rock shows a weak tectonic overprint, it is evident that the granitoid target rocks underwent catastrophic and pervasive shear faulting.

At 1,220 to 1,316 m.b.s.f., the target rock is strongly distorted and brecciated, and fragments of it are marginally resorbed and found in melt rock (Fig. 5a, b). Conversely, zones of brecciated target rock host elongated, and frequently wispy, melt-rock fragments, reminiscent of suevite (Fig. 5c, d). Where in contact with target rock fragments, the melt rock underwent large ductile strains, as is clear from the highly stretched granitoid fragments contained in the melt rock (Fig. 5e). Overall, the melt rock is spatially associated with the highest-strained target rocks, indicated by breccia of thicknesses of decimetres to metres. The presence of exotic fragments (Fig. 5f)—consisting of gneiss, mafic igneous rock and various mylonites—in the melt rock excludes an in situ frictional melt origin for the melt rock. Breccia zones are substantially thicker and show a larger range in sizes and shapes of fragments than cataclasite and ultra-cataclasite zones in target rock outside this particular depth interval. The differences in thickness and fragment size between these breccia and the cataclasite zones indicate different fragmentation mechanisms and/or fragmentation at different times during the cratering process. Finally, the spatial density of ductile band structures is maximal within

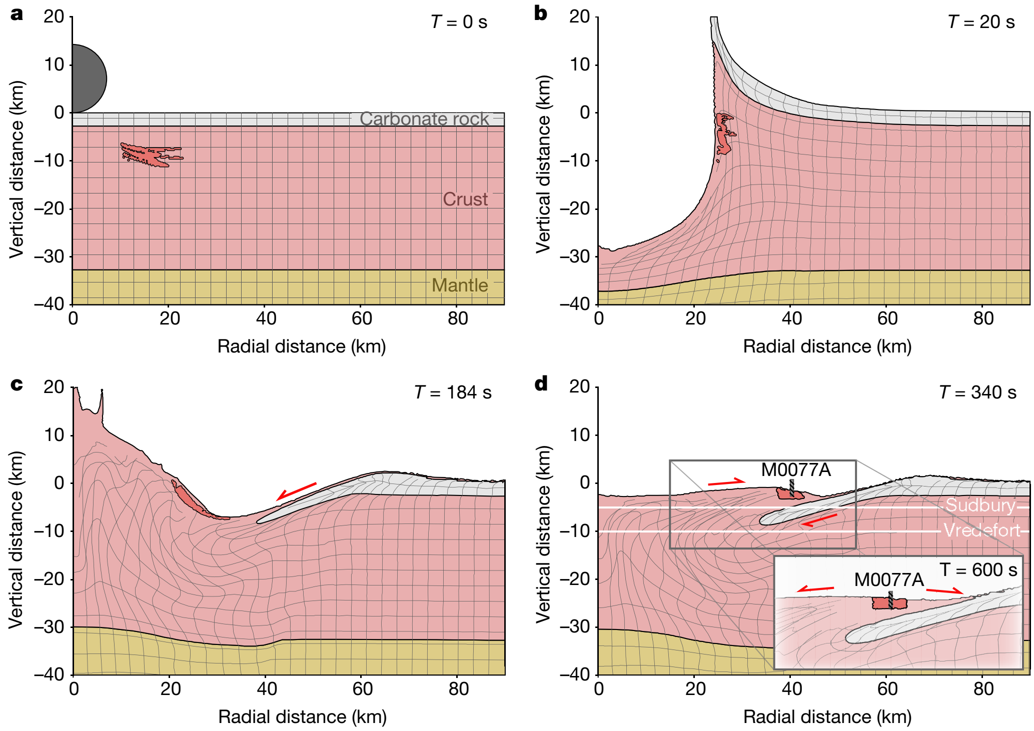


Fig. 2 | Modelled formation of the Chicxulub impact structure. The mechanism is based on numerical modelling of peak-ring crater formation[4,23,24,34]. A grid of tracer particles is shown to highlight the sub-crater deformation. Dark red area of crust in each panel tracks the material that eventually forms the peak ring. *T* denotes time in seconds after impact. Red half-arrows indicate the direction of major shear displacements relative to adjacent material. **a**, Undisturbed configuration of model lithosphere before impact. **b**, Cratering starts by shock-wave-induced, crustal-scale excavation of a bowl-shaped transient cavity. **c**, Gravitational instability of the transient cavity causes uplift of the crater centre and concomitant inward slumping of the cavity wall. **d**, Collapse and radial outward displacement of uplifted material over inward-slumped cavity wall segments followed by gravitational settling of the peak ring (inset) characterize the terminal phase of modelled crater modification. White lines indicate the approximate current erosion levels of the Sudbury and Vredefort impact structures.

this depth interval (Fig. 3). Brittle–ductile band structures occur predominantly in mechanically and thermally weakened target and melt rock and form ductile shear zones (Fig. 5f), shear bands with C–S fabric geometry (Fig. 5g)[31] and crenulated mineral fabrics (Fig. 4h).

Chronology of deformation mechanisms

Most importantly, it is possible to determine the relative timing of the various deformation mechanisms. Zones of (ultra-)cataclasite truncate the jigsaw fragment geometry of pervasively fractured target rock (Fig. 4a, b). Shear faults, in turn, consistently offset cataclasite and ultra-cataclasite zones (Figs. 4c and 5a, b). Target rock fragments in melt rock are sporadically striated and host cataclasite zones[28]; whereas melt rock matrices are devoid of shear faults. Cataclasite and melt rock are found in tension fractures (Fig. 5h), which, to some extent, formed from shear faults. Brittle–ductile band structures displace zones of cataclasite, crenulated foliation surfaces and the contacts of target rock with cataclasite and melt rock (Figs. 4g, h and 5f, g). In summary, pervasive fracturing of target rock was followed, respectively, by formation of (ultra-)cataclasite zones, shear faulting, emplacement of cataclasite and impact melt into dilatant fractures and formation of ductile band structures.

Deformation mechanisms and cratering stages

During the various cratering stages, deformation kinematics and states of stress of the target rock differ profoundly (Fig. 2). Therefore, distinct deformation mechanisms recognized in the target rock may well relate to individual cratering stages denoted in terms of time after impact. Shock and decompression causes irreversible plastic deformation and imparts to the shocked rocks a divergent outward velocity field, which forms the transient cavity. This velocity field causes wall-parallel extension and perpendicular shortening of the target rock (Fig. 2b). Rock deformation at upper-crustal pressures and depths, which is where peak-ring materials are derived from, is accommodated by fracturing. We therefore attribute pervasive fracturing, which preceded the other deformation mechanisms, to shock loading, decompression and transient cavity growth (time after impact $T < 30$ s).

After the transient cavity forms (Fig. 2b), gravitational collapse modifies the crater shape until the final crater morphology is reached (Supplementary Information). During initial collapse, the peak-ring material motion transitions from outward and divergent excavation flow to inward and convergent rock flow towards the crater centre. This inward movement leads to the incorporation of peak-ring material onto the flank of a central uplift (Fig. 2c). During this stage of cratering, peak-ring materials experience several distinct stress states (Extended Data Fig. 2). Planar zones of cataclasite and (ultra-)cataclasite are plausible candidates for accommodating the deformation of pre-fractured target rock during this cratering stage ($20\ \mathrm{s} < T < 150\ \mathrm{s}$).

During build-up of the central uplift ($20\ \mathrm{s} < T < 100\ \mathrm{s}$), the pressure on the peak-ring material increases (Extended Data Fig. 2). This increase inevitably closes asperities within the fractured rock and thus increases the internal friction of the target rock and normal stresses on faults. The central uplift eventually over-heightens and becomes gravitationally unstable, causing downwards and radial-outward collapse ($160\ \mathrm{s} < T < 300\ \mathrm{s}$). In this motion, collapsed material piles up to form the peak ring, which is thrust over the inwardly slumped transient cavity rim (Fig. 2d, Supplementary Information). Collectively, the increased pressure, combined with the reversal of the material displacement field as the central uplift transitions from motion upwards to outwards and downwards during collapse accounts for the observed transition from localized cataclastic flow to shear faulting during this stage of cratering.

As the melt rock occurrences within the target rock are devoid of shear faults, melt emplacement must occur at the end of peak-ring

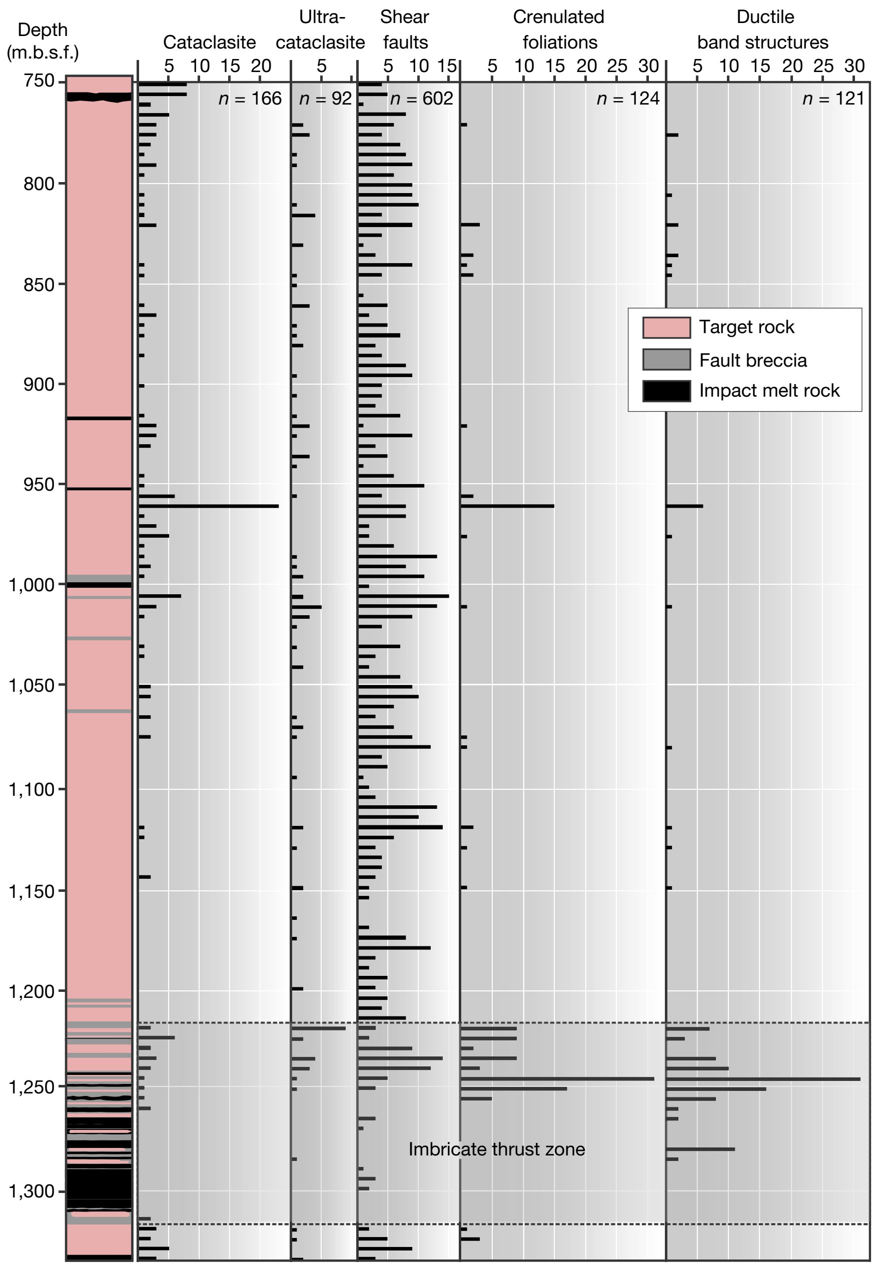


Fig. 3 | Spatial distribution of major lithological units and deformation structures in target rock of M0077A drill core. *n* is number of observations. We note the strong spatial correlation of increased numbers of (ultra-)cataclasite zones, crenulated foliations and ductile band structures below 1,220 m.b.s.f.

formation (250 s < T < 600 s). Subsequent deformation is evident from the ductile band structures displacing contacts between the target and melt rock, zones of cataclasite and mineral foliations. Band orientation, the sense of displaced layers and fabric asymmetry, as displayed by sigmoidal foliation planes and cataclasite zone boundaries, consistently indicate band formation through normal faults (Figs. 4g, h and 5f, g). Respective vertical shortening and horizontal extension is consistent with gravitational spreading of the topographically elevated peak ring and signifies the final stage of crater modification (inset in Fig. 2d).

Weakening mechanisms

The recognition of distinct deformation mechanisms corresponding to the various stages of the cratering process is of fundamental importance in comprehending the mechanics of large-scale impact cratering. Initial pervasive grain-scale fracturing causes a profound loss of cohesion in target rocks at the onset of, and during, transient cavity growth. During cavity modification, strain is localized progressively through formation of cataclasite zones, ultra-cataclasite zones, shear faults, and finally deformation on fault zones with impact-melt-bearing fault breccias. Progressive strain localization is evidence of the incremental regaining of shear and cohesive strength in the target rock, as crater modification proceeds. It has been proposed that crater collapse is facilitated by the self-lubrication of faults by frictional melts[18]. We did not, however, uncover any evidence for friction-generated melt rock in the recovered target rock from the peak ring at Chicxulub. Hence, dynamic weakening of faults, if important, appears to require a mechanism other than shear heating.

Shock compression and dilation during initial impact caused wholesale intra-crystalline damage (Fig. 4e, f). Thereafter, dynamic fracturing induced by the passage of the shock and rarefaction waves and transient cavity growth led to loss in cohesion and shear strength. The presence of pervasively fractured target rock with preserved microscopic jigsaw fragment patterns and uniform orientation of pre-impact dykes (Extended Data Fig. 1) indicate that target rock above 1,220 m.b.s.f.

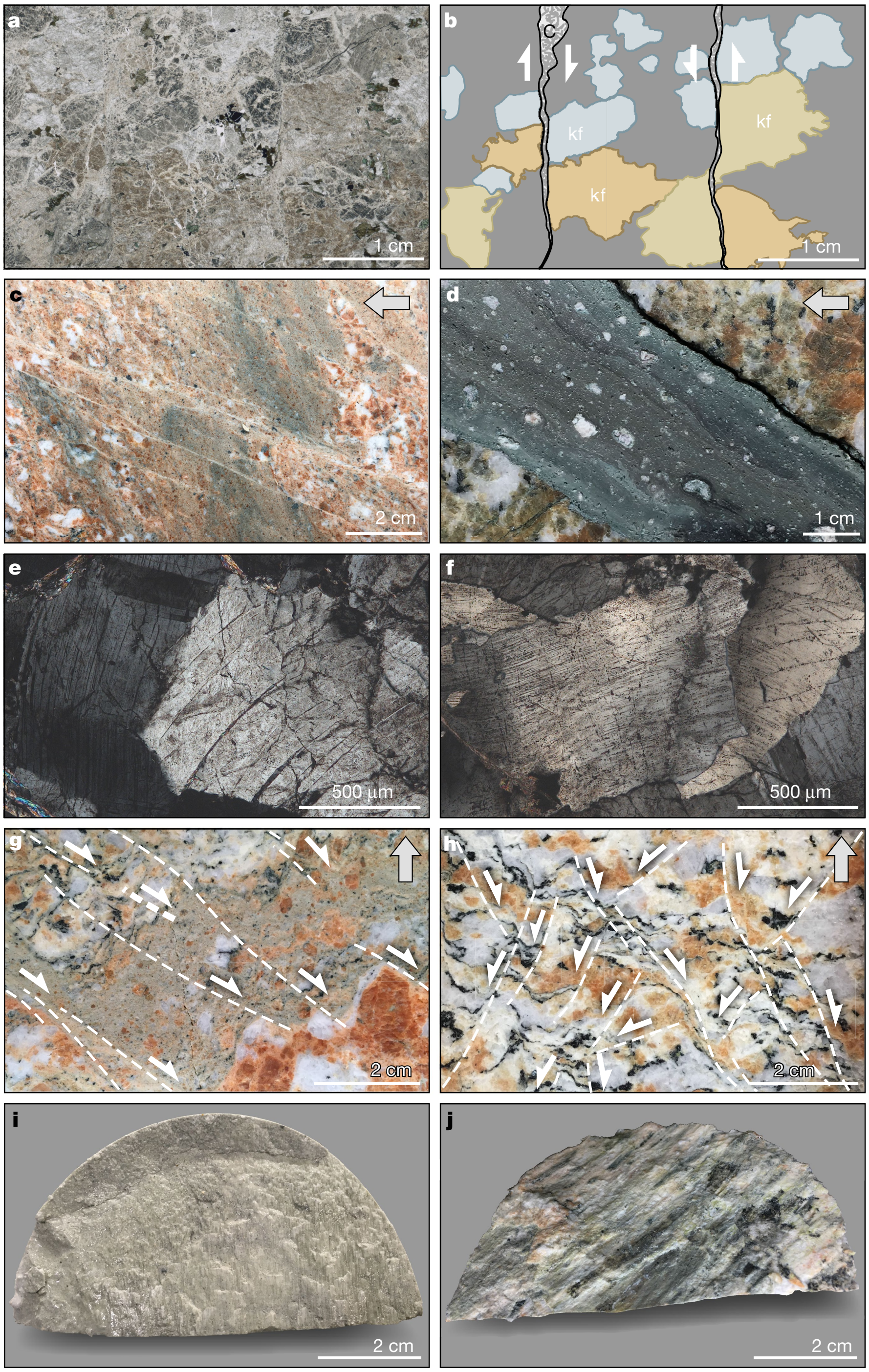


Fig. 4 | Deformation structures in target rock at site M0077A. Arrow indicates the direction of the top of the drill core. Half-arrows indicate the sense of displacement on discontinuities. **a**, Photomicrograph in plane-polarized light showing pervasive cataclasite of granitoid target rock (core 122-3, 820 m.b.s.f.). **b**, Line drawing of **a** showing alkali-feldspar (kf) displaced on cataclasite zone (**c**). **c**, Cataclasite zones displaced on shear faults (core 301-1, 1,326.45–1,326.57 m.b.s.f.). **d**, Flow-foliated ultra-cataclasite (core 215-2, 1,065.85–1,065.94 m.b.s.f.). **e**, Photomicrograph in cross-polarized light showing distorted twin lamellae in plagioclase (core 129-1, 831.38–831.40 m.b.s.f.). **f**, Photomicrograph in cross-polarized light showing distorted quartz with planar deformation features (core 129-1, 831.38–831.40 m.b.s.f.). **g**, Cataclasite zone segmented by normal faults (core 172-2, 956.41–956.45 m.b.s.f.). **h**, Crenulated layering in granitoid rock (core 122-1, 817.61–817.66 m.b.s.f.). **i**, Striated shear fault in carbonate rock. **j**, Striated shear fault in granitoid target rock (core 154-1, 894.19 m.b.s.f.).

behaved largely as a structurally coherent rock mass. The implication of small displacements across the entire rock mass is consistent with macroscopic deformation of an acoustically fluidized rock mass[19,20].

Structural observations from the peak-ring target rocks of Chicxulub are generally consistent with acoustic fluidization as the dominant weakening mechanism and offer insight for the refinement of future impact simulations. Acoustic fluidization entails target rock blocks undergoing pressure oscillations around the ambient lithostatic stress[4,7,19,20]. During pressure lows, blocks have reduced normal stresses between them, drastically reducing frictional resistance at block boundaries during periodic rock flow. During pressure highs, blocks are compressed, locally increasing the frictional resistance of the deforming rock mass. Cataclasite zones seem likely to be where the sheared block boundaries serve as contact strain zones during oscillation of target

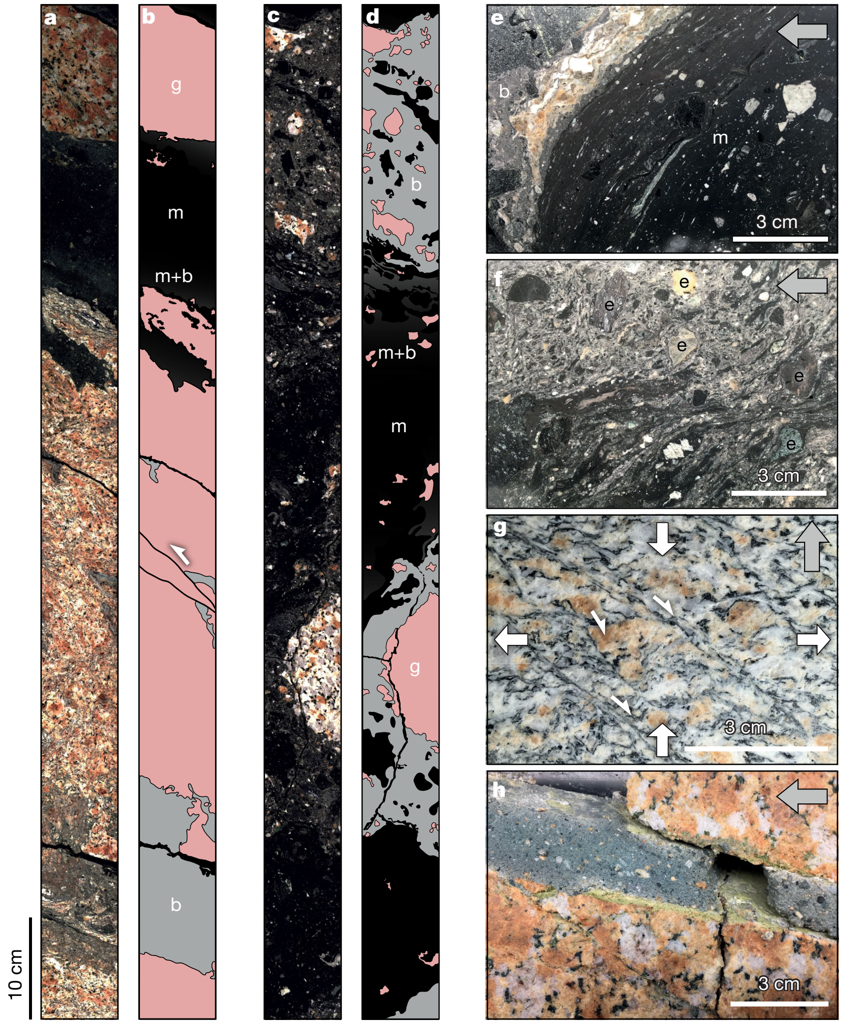


Fig. 5 | Images illustrating rock types found between 1,220 and 1,316 m.b.s.f. (g, highly distorted granitoid rock; m, impact melt rock; b, fault breccia; e, exotic fragments.) **a**, Line scan of core 265-2 (1,216.36–1,217.45 m.b.s.f.), showing highly distorted and brecciated target rock mingled within melt rock. We note halos of mingled melt rock and fault breccia at the margins of granitoid rocks as well as shear faults (half-arrow) displacing thin zones of ultra-cataclasite. **b**, Line drawing of **a**. **c**, Line scan of core 285-1 (1,277.24–1,278.25 m.b.s.f.), displaying mingling of impact melt rock and fault breccia, notably near the granitoid fragment. We note melt rock fragments within fault breccia. **d**, Line drawing of **c**. **e**, Melt rock in contact with fault breccia. We note the gradient in contact strain, evident from the stretched target rock fragments in melt rock (core 303-3, 1,334.24–1,334.35 m.b.s.f.). **f**, Ductile shear zone in mingled impact melt rock and fault breccia containing exotic fragments (core 289-1, 1,289.75–1,289.87 m.b.s.f.). **g**, C–S fabric geometry in granitoid indicated by displaced planar mineral fabric in granitoid target rock (half arrows) amounting to vertical shortening and horizontal extension (white arrows) (core 273-2, 1,241.26–1,241.31 m.b.s.f.). **h**, Cataclasite entrained in dilatant fracture (core 262-1, 1,207.45–1,207.56 m.b.s.f.).

rock blocks. Continued cataclasis, resulting in flow-foliated ultra-cataclasite, heralds an increase in shear strain of the rock mass and waning acoustic fluidization. While in motion, continued comminution in (ultra-)cataclasite zones may generate additional acoustic energy and prolong cataclastic flow[20].

A critical parameter in the acoustic fluidization model is the dominant wavelength of pressure vibrations[19], which controls both the viscosity of the acoustically fluidized rock mass and the timescale for the decay of vibrations. The 'block model' of acoustic fluidization is employed in most Chicxulub-scale impact simulations[4,23,24], such as the one reproduced in Fig. 2. The block model supposes that the subcrater rock mass is dominated by blocks of a characteristic size that oscillate within a surrounding mass of breccia with a single vibrational wavelength (and period) that is directly proportional to the block size[32]. The block model parameters employed in Chicxulub impact simulations imply a block size of about 100–500 m (depending on the assumed acoustic energy dissipation factor Q) and an oscillation frequency of a few hertz. This prediction is consistent with the entire approximately 450-m granite sequence above the imbricate thrust zone representing a single 'block' (Fig. 3).

On the other hand, if the cataclasite zones observed in the Chicxulub peak-ring drill core represent oscillating-block boundaries as we propose, their average spacing (Fig. 3) of about 3.5 m (2.3 m including ultra-cataclasite zones) would imply a much smaller block size, shorter dominant vibrational wavelength and higher vibrational frequency[19,20]. This would imply rapid evolution of the acoustic wave field during collapse of the crater, which is not predicted by the current block model implementation. High-frequency vibrations sustained for the duration of crater collapse, however, could be explained by the efficient regeneration of acoustic energy during the cratering process, which is neglected in the block model. Effective regeneration of vibrations in a rapidly shearing rock mass is consistent with findings from discrete-element models of acoustic fluidization in landslides[33]. Alternatively, the acoustic wave field may evolve by progressive lengthening of the dominant vibrational wavelength during cratering as higher-frequency vibrations dissipate sooner. In this case, the effective block size could increase during crater formation from a few metres at the beginning of modification, when the first cataclasite zones are likely to have formed ($20\text{ s} < T < 60\text{ s}$), to a few hundred metres by the end of peak-ring emplacement ($T < 600$ s).

A progressive waning of the acoustic wavefield in which slip events, facilitated by negative pressure excursions, become less frequent and more widely spaced is consistent with the temporal evolution of deformation observed in the drill core. This evolution suggests a progression from distributed, small-displacement deformation along closely spaced faults early in the cratering process to more localized, larger-displacement deformation along widely spaced slip surfaces later. Acoustic fluidization is, therefore, interpreted to halt at the onset of shear faulting, as target rock blocks cease to oscillate and the bulk rock mass regains internal friction and, thus, shear strength. Whether this cessation of acoustic fluidization occurs during the final emplacement of the peak ring (as suggested by current numerical simulations; Fig. 2d) or earlier, during the formation of the central uplift, is unclear. In the latter scenario, the outward collapse of the central uplift and thrusting of peak-ring rocks onto the transient cavity rim would have occurred after the rocks regained most of their large-scale static strength. In this case, the late stages of collapse could have been facilitated by large faults, lubricated by entrained impact melt.

Peak-ring formation

Modelling suggests that the target rock forming the peak ring resided at a depth[24] of about 10 km, before impact, and was entrained into a central uplift before being thrust outward over inward slumped transient cavity wall segments (Fig. 2). From the modelled cratering flow (Supplementary Information), it is conceivable that individual target rock blocks may over-thrust portions of impact melt, notably where the peak ring develops. Impact melt may then become sandwiched between quasi-coherent target rock masses. Hence, impact melt in large craters may be present not only as ponded liquids at the surface, but also as melt bodies or sheets entrained and trapped in target rock thrust zones at depth.

Structural and lithological characteristics of the rocks at depths between 1,220 and 1,316 m.b.s.f. are consistent with impact melt entrained in a prominent imbricate thrust zone (Fig. 3). Respective characteristics include: (1) the concentration of high strains in target rock and melt rock (Fig. 5a, e), (2) the strongly distorted target rock slivers mingled with melt rock and breccia, interpreted as fault breccia (Fig. 5a–d, f), (3) the occurrence of melt rock fragments in fault breccia (Fig. 5c, d, f), and (4) fragment lithologies not present in the adjacent target rock[28]. Given that in situ frictional melting is excluded for the origin of the melt rock, formation of this rock by shock-induced melting and subsequent entrainment during peak-ring formation appears to be the more plausible explanation. Specifically, we propose that the target rock mass above 1,220 m.b.s.f. over-thrust and buried the impact melt overlying the deeper target rock, which is now found below 1,316 m.b.s.f. Impact melt rock in contact with brecciated target rock displays large ductile strains (Fig. 5e) and indicates rapid cooling (quenching) and solidification of the impact melt during thrusting. In summary, imbricate thrusting (stacking) of target rock masses[34] contributed to the high topography of the peak ring. A prerequisite for thrusting is the regaining of shear strength in the target rock by the time of the formation of peak-ring topography.

Consequences of dynamic weakening

Examination of the deformation mechanisms of the target rocks underlying the peak ring at Chicxulub has provided unprecedented evidence for the physical mechanisms responsible for weakening and the regained strength of target rock during large-scale impact cratering. Results are strongly supportive of the dynamic collapse model (Fig. 2, Supplementary Information) of peak-ring formation and of acoustic fluidization as the dominant mechanism driving crater modification. The transition in deformation style from distributed cataclastic flow to localized shear-faulting and the progressive increase in fault spacing illuminates the waning of acoustic fluidization and the target regaining sufficient strength to support the topography of the peak ring. Dynamic weakening of faults or regeneration of acoustic energy may have an important role in this final phase of peak-ring formation. Incorporating this insight into future numerical impact simulations will aid in the design of higher-fidelity models of large-scale impact cratering.

In particular, we regard (ultra-)cataclasite zones, serving as contact strain zones of oscillating target rock blocks, as the physical manifestation of pressure fluctuations. If so, the estimated average size of coherent target rock blocks within the Chicxulub peak ring is one to two orders of magnitudes smaller than observed in the central uplifts of smaller terrestrial complex craters[35–37]. This may imply efficient regeneration of pressure fluctuations during transient cavity collapse and modification or a growth in vibrational wavelength as the wavefield evolves. In either case, central peaks of smaller impact structures may be preserved because fluidization ceased early in the gravitational collapse process. By contrast, peak rings in peak-ring craters and multi-ring basins form because acoustic fluidization is sustained through the formation and collapse of an overheightened central uplift.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0607-z.

Received: 28 February 2018; Accepted: 15 August 2018;
Published online 24 October 2018.

1. Croft, S. K. The modification stage of basin formation: conditions of ring formation. *Geochim. Cosmochim. Acta* **12A**, 227–257 (1981).
2. Grieve, R. A. F., Robertson, P. B. & Dence, M. R. Constraints on the formation of ring impact structures, based on terrestrial data. *Geochim. Cosmochim. Acta* **12A**, 37–57 (1981).
3. Neumann, G. A. et al. Lunar impact basins revealed by gravity recovery and interior laboratory measurements. *Sci. Adv*. **1**, e1500852 (2015).
4. Ivanov, B. A. Numerical modelling of the largest terrestrial meteorite craters. *Sol. Syst. Res*. **39**, 381–409 (2005).
5. Melosh, H. J. & Ivanov, B. A. Impact crater collapse. *Annu. Rev. Earth Planet. Sci*. **27**, 385–415 (1999).
6. O'Keefe, J. D. & Ahrens, T. J. Planetary cratering mechanics. *J. Geophys. Res*. **98**, 17011–17028 (1993).
7. Wünnemann, K. & Ivanov, B. A. Numerical modelling of the impact crater depth-diameter dependence in an acoustically fluidized target. *Planet. Space Sci*. **51**, 831–845 (2003).
8. Grady, D. E. & Kipp, M. E. in *Fracture Mechanics of Rock* (ed. Atkinson, B. K.) 429–475 (Academic Press, London, 1987).
9. Fujiwara, A., Kamimoto, G. & Tsukamoto, A. Destruction of basaltic bodies by high-velocity impact. *Icarus* **31**, 277–288 (1977).
10. Ahrens, T. J. & Rubin, A. M. Impact-induced tensional failure in rock. *J. Geophys. Res*. **98**, 1185–1203 (1993).
11. Buhl, E. et al. Particle size distribution and strain rate attenuation in hypervelocity impact and shock recovery experiments. *J. Struct. Geol*. **56**, 20–33 (2013).
12. Collins, G. S. Numerical simulations of impact crater formation with dilatancy. *J. Geophys. Res*. **119**, 2600–2619 (2014).
13. Melosh, H. J., Ryan, E. V. & Asphaug, E. Dynamic fragmentation in impacts: hydrocode simulation of laboratory impacts. *J. Geophys. Res*. **97**, 14,735–14,759 (1992).
14. Kenkmann, T. Folding within seconds. *Geology* **30**, 231–234 (2002).
15. Kenkmann, T. Dike formation, cataclastic flow, and rock fluidization during impact cratering: an example from the Upheaval Dome structure, Utah. *Earth Planet. Sci. Lett*. **214**, 43–58 (2003).
16. Collins, G. S., Melosh, H. J. & Ivanov, B. A. Modeling damage and deformation in impact simulations. *Meteorit. Planet. Sci*. **39**, 217–231 (2004).
17. Senft, L. E. & Stewart, S. T. Dynamic fault weakening and the formation of large impact craters. *Earth Planet. Sci. Lett*. **287**, 471–482 (2009).
18. Spray, J. G. Superfaults. *Geology* **25**, 579–582 (1997).
19. Melosh, H. J. Acoustic fluidization: a new geological process? *J. Geophys. Res*. **84**, 7513–7520 (1979).
20. Melosh, H. J. Dynamical weakening of faults by acoustic fluidization. *Nature* **379**, 601–606 (1996).
21. Grieve, R. A. F. & Therriault, A. Vredefort, Sudbury, Chicxulub: three of a kind? *Annu. Rev. Earth Planet. Sci*. **28**, 305–338 (2000).
22. Grieve, R. A. F., Reimold, W. U., Morgan, J., Riller, U. & Pilkington, M. Observations and interpretations at Vredefort, Sudbury and Chicxulub: towards a composite kinematic model of terrestrial impact basin formation. *Meteorit. Planet. Sci*. **43**, 855–882 (2008).
23. Collins, G. S., Melosh, H. J., Morgan, J. V. & Warner, M. R. Hydrocode simulations of Chicxulub crater collapse and peak-ring formation. *Icarus* **157**, 24–33 (2002).
24. Morgan, J. V. et al. The formation of peak rings in large impact craters. *Science* **354**, 878–882 (2016).

25. Morgan, J. V., Warner, M. R., Collins, G. S., Melosh, H. J. & Christeson, G. L. Peak-ring formation in large impact craters: geophysical constraints from Chicxulub. *Earth Planet. Sci. Lett.* **183**, 347–354 (2000).
26. Morgan, J. V. et al. Full waveform tomographic images of the peak ring at the Chicxulub impact crater. *J. Geophys. Res.* **116**, B06303 (2011).
27. Gulick, S. P. S. et al. Importance of pre-impact crustal structure for the asymmetry of the Chicxulub impact crater. *Nat. Geosci*. **1**, 131–135 (2008).
28. Morgan, J. V. et al. Chicxulub: drilling the K-Pg impact crater. In *Proceedings of the International Ocean Discovery Program 364* https://doi.org/10.14379/iodp.proc.364.2017 (International Ocean Discovery Program, College Station, 2017).
29. Christeson, G. L. et al. Extraordinary rocks from the peak ring of the Chicxulub impact crater: P-wave velocity, density, and porosity measurements from IODP/ICDP Expedition 364. *Earth Planet. Sci. Lett.* **495**, 1–11 (2018).
30. Petit, J. P. Criteria for the sense of movement on fault surfaces in brittle rocks. *J. Struct. Geol*. **9**, 597–608 (1987).
31. Berthé, D., Choukroune, P. & Jegouzo, P. Orthogneiss, mylonite and non-coaxial deformation of granites: the example of the South Armorican Shear Zone. *J. Struct. Geol*. **1**, 31–42 (1979).
32. Ivanov, B. A. & Artemieva, N. A. in *Catastrophic Events and Mass Extinctions: Impact and Beyond* (eds C. Koeberl, C. & MacLeod, K. G.) Geological Society of America Special Paper **356**, 619–630 (GSA, 2002).
33. Johnson, B. C., Campbell, C. S. & Melosh, H. J. The reduction of friction in long runout landslides as an emergent phenomenon. *J. Geophys. Res*. **121**, 881–889 (2016).
34. Kring, D. A., Kramer, G. Y., Collins, G. S., Potter, R. W. K. & Chandnani, M. Peak-ring structure and kinematics from a multi-disciplinary study of the Schrödinger impact basin. *Nat. Commun*. **7**, 13161 (2016).
35. Ivanov, B. A., Kocharyan, G. G. & Kostuchenko, V. N. Puchezh-Katunki impact crater: preliminary data on recovered core block structure. In *Proc. 27th Lunar and Planetary Science Conf*. 589–590, https://www.lpi.usra.edu/meetings/lpsc1996/pdf/1295.pdf (1996).
36. Kenkmann, T., Collins, G. S. & Wünnemann, K. in *Impact Cratering: Processes and Products* (eds Osinski, G. R. & Pierazzo, E.) 60–75 (John Wiley & Sons, Chichester, 2013).
37. Rae, A. S. P., Collins, G. S., Grieve, R. A. F., Osinski, G. R. & Morgan, J. V. Complex crater formation: insights from combining observations of shock pressure distribution with numerical models at the West Clearwater Lake impact structure. *Meteorit. Planet. Sci*. **52**, 1330–1350 (2017).

Acknowledgements This work was supported by the Priority Programs 527 and 1006 of the German Science Foundation (grants Ri 916/16-1 and PO 1815/2-1), National Science Foundation grants (OCE-1737351, OCE-1450528 and OCE-1736826), and Natural Environment Research Council (grants NE/P011195/1 and NE/P005217/1). The Chicxulub drilling expedition was funded by the European Consortium for Ocean Research Drilling (ECORD) and the IODP as Expedition 364 with co-funding from the ICDP. The Yucatan State Government and Universidad Nacional Autónoma de México (UNAM) provided logistical support. This research used samples and data provided by IODP. Samples can be requested at http://web.iodp.tamu.edu/sdrm. We are grateful for assistance from the staff of the IODP Core Repository in Bremen, Germany, during the Onshore Science Party. We thank B. Ivanov and C. Koeberl for constructive reviews and S. Teuber for assistance in figure preparation. This is UTIG contribution number 3,278.

Reviewer information *Nature* thanks B. Ivanov and C. Koeberl for their contribution to the peer review of this work.

Author contributions U.R., M.H.P., A.S.P.R., J.V.M., S.P.S.G. and R.A.F.G. conceived the study. All authors participated in sampling and data collection offshore and/or onshore during IODP-ICDP Expedition 364, interpretation of the data as well as writing and/or editing of the manuscript. U.R. provided the first draft of the manuscript. U.R. and F.M.S. acquired structural data from line scans. J.L. and A.D. provided the downhole orientation data. A.S.P.R and G.S.C. performed and analysed the numerical models; G.S.C., A.S.P.R. and H.J.M. contributed the discussion on the implications for acoustic fluidization.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0607-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0607-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to U.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

IODP–ICDP Expedition 364 Science Party
Joanna V. Morgan[3], Sean P. S. Gulick[6,7], Sophie L. Green[11], Johanna Lofi[8], Elise Chenot[12], Gail L. Christeson[6], Philippe Claeys[13], Charles S. Cockell[14], Marco J. L. Coolen[15], Ludovic Ferrière[16], Catalina Gebhardt[17], Kazuhisa Goto[18], Heather Jones[19], David A. Kring[9], Long Xiao[20], Christopher M. Lowery[6,7], Rubén Ocampo-Torres[21], Ligia Perez-Cruz[22], Annemarie E. Pickersgill[23,24], Michael H. Poelchau[25], Auriol S. P. Rae[26], Cornelia Rasmussen[6,7], Mario Rebolledo-Vieyra[27], Ulrich Riller[1], Honami Sato[28], Jan Smit[29], Sonia M. Tikoo-Schantz[30], Naotaka Tomioka[31], Michael T. Whalen[32], Axel Wittmann[33], Kosei Yamaguchi[34,35], Jaime Urrutia Fucugauchi[22] & Timothy J. Bralower[19]

[11]British Geological Survey, The Lyell Centre, Research Avenue South, Edinburgh, UK. [12]Université de Bourgogne-CNRS, Biogeosciences Laboratory, Dijon, France. [13]Analytical, Environmental and Geochemistry (AMGC), Vrije Universiteit Brussel (VUB), Brussels, Belgium. [14]School of Physics and Astronomy, UK Center for Astrobiology, University of Edinburgh, Edinburgh, UK. [15]Western Australia Organic and Isotope Geochemistry Centre, School of Earth and Planetary Sciences, Curtin University, Bentley, Western Australia, Australia. [16]Natural History Museum, Vienna, Austria. [17]Alfred Wegener Institute Helmholtz Centre of Polar and Marine Research, Bremerhaven, Germany. [18]International Research Institute of Disaster Science, Tohoku University, Sendai, Japan. [19]Pennsylvania State University, University Park, PA, USA. [20]China University of Geosciences (Wuhan), School of Earth Sciences, Planetary Science Institute, Wuhan, China. [21]National Center of Scientific Research (CNRS), Groupe de Physico-Chimie de l'Atmosphère, Institut de Chimie et Procédés pour l'Energie, l'Environnement et la Santé ICPEES, Université de Strasbourg, Strasbourg, France. [22]Instituto de Geofísica, Universidad Nacional Autónoma De México, México City, Mexico. [23]School of Geographical and Earth Sciences, University of Glasgow, Glasgow, UK. [24]Argon Isotope Facility, Scottish Universities Environmental Research Centre (SUERC), East Kilbride, UK. [25]Department of Geology, University of Freiburg, Freiburg, Germany. [26]Department of Earth Science and Engineering, Imperial College London, London, UK. [27]Unidad de Ciencias del Agua, Mérida, Mexico. [28]Japan Agency for Marine-Earth Science and Technology, Yokosuka, Japan. [29]Faculty of Earth and Life Sciences, Amsterdam, The Netherlands. [30]Earth and Planetary Sciences, Rutgers University—New Brunswick, Piscataway, NJ, USA. [31]Japan Agency for Marine-Earth Science and Technology, Kochi Institute for Core Sample Research, Kochi, Japan. [32]Department of Geosciences, University of Alaska Fairbanks, Fairbanks, AK, USA. [33]Eyring Materials Center, Arizona State University, Tempe, AZ, USA. [34]Department of Chemistry, Tohu University, Funabashi, Japan. [35]NASA Astrobiology Institute, Mountain View, CA, USA.

METHODS

Acquisition of structural data from drill core. In addition to the methods employed for visual appraisal as well as meso- and microstructural analyses of the drill core during the Onshore Science Party[38], the following analyses were conducted. On the basis of a detailed examination of drill core line-scans, the occurrence of cataclasite zones, ultra-cataclasite zones, crenulated foliations and ductile band structures was recorded with depth. Only zones of (ultra-)cataclasite displaying a thickness of 1 cm and larger were recorded. Distinction between the two types of cataclasite is based on grain size, the presence of flow foliation and the fragment-size distribution. Overall, ultra-cataclasite appears darker than cataclasite. Mesoscopic shear faults displaying slip lineations and slip sense were identified by carefully removing core sections from the liners. Statistical analysis of the spatial occurrence of the structures was conducted with Microsoft Excel (see Source Data for Fig. 3).

Microstructural analysis. Polished thin sections of thickness 25 μm were produced from selected target rock samples at the Institute of Mineralogy and Petrography of the University of Hamburg, Germany. Microscopic inspection of thin sections was conducted using a Zeiss Axio Scope.A1 polarization microscope and attached high-resolution digital camera AxioCam MRc Rev. 3 FireWire.

Borehole imaging of planar structures. During Expedition 364, both optical and acoustic borehole images of the borehole walls were acquired[38]. Post-acquisition processing and analysis allowed manual picking of the planar structural discontinuities, corresponding to pre-impact igneous sheet intrusions, and determination of their orientation. Orientations have not been corrected from borehole deviation, which departs less than 4° from the vertical. For visualization and processing of borehole images, the ALT WellCAD (https://www.alt.lu/products-wellcad/) software package was used. For analysis of orientation of pre-impact sheet intrusions the software package Tectonics FP version 1.6 was used[39].

Numerical modelling. To aid interpretation of the drill core data, we reproduced and reprocessed the numerical simulation of the Chicxulub impact[24], which was in turn based on previous Chicxulub impact simulations that produced a good match to geological and geophysical constraints[4,23,32,40]. The impactor parameters of the model were: diameter 14 km, velocity 12 km s^{-1}, density 2,630 kg m^{-3}. A vertical incidence impact angle was enforced by the cylindrical geometry of the two-dimensional model. A spatial resolution of 200 m was used, corresponding to 35 cells across the impactor radius. A simplified target structure was used of 3 km (carbonate) cover rocks and 30 km (granite) crust overlying (dunite) mantle. The simulation duration was 600 s of model time. We refer to ref. [24] for a full description of the modelling approach, including a comprehensive list of model parameters.

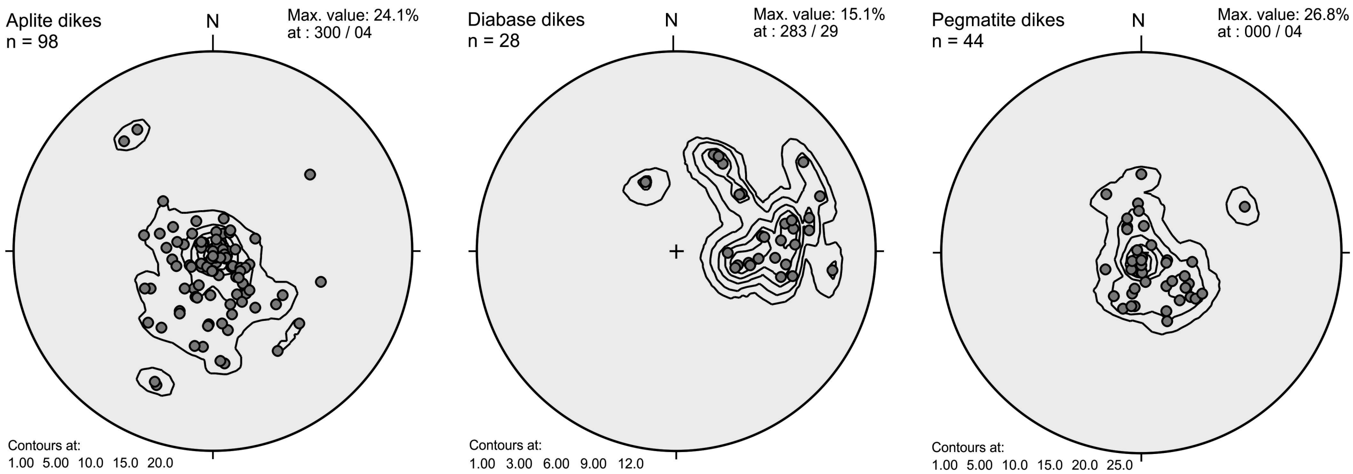
Simulations were processed to examine the motion and pressure of peak-ring materials (Fig. 2a–d, Extended Data Fig. 2, Supplementary Video). Lagrangian tracer particles employed in the numerical method allow the history of material that ends up within the peak ring to be recorded and interrogated. Ref. [24] used tracer particles to illustrate the peak pressure and provenance of the peak-ring materials, as well as its motion during crater formation. Here, we identified a subset of 100 tracer particles within the same peak-ring material, initially located within a square (2 km × 2 km) cross-section at a depth of 10 km and a radius of 16 km (see Supplementary Information, $T = 0$). The Supplementary Video shows the motion of these tracers during cratering in both the fixed simulation reference frame (main image) and in a Lagrangian reference frame, centred on the average location of the 100 tracers (inset). The inset image gives a qualitative sense of the internal deformation of the peak-ring materials and highlights the deformation kinematics of peak-ring material during cratering.

Additionally, we analysed the pressure recorded by each tracer (circles) within the same volume, as well as the average pressure (solid line), as a function of time during the simulation (Extended Data Fig. 2). After the brief passage of the shock wave ($P > 10$ GPa; $T < 5$ s), the pressure in the peak-ring materials rises from 10–20 MPa to 50–100 MPa between about 100 s and about 250 s, before returning back to 10–20 MPa. Thus, the inward collapse of the peak-ring materials towards the central uplift and the subsequent outward collapse are associated with elevated pressures, above the ultimate overburden pressure in the peak-ring materials at their final location. We note that pressure waves caused by shockwave reflections from the numerical domain boundaries, which would not be present in reality, are superimposed on the pressure–time signal after about 130 s. While these complicate interpretation, the elevated pressure for the two minutes of central uplift formation and collapse is a robust outcome of the model that is insensitive to the location of the domain boundary.

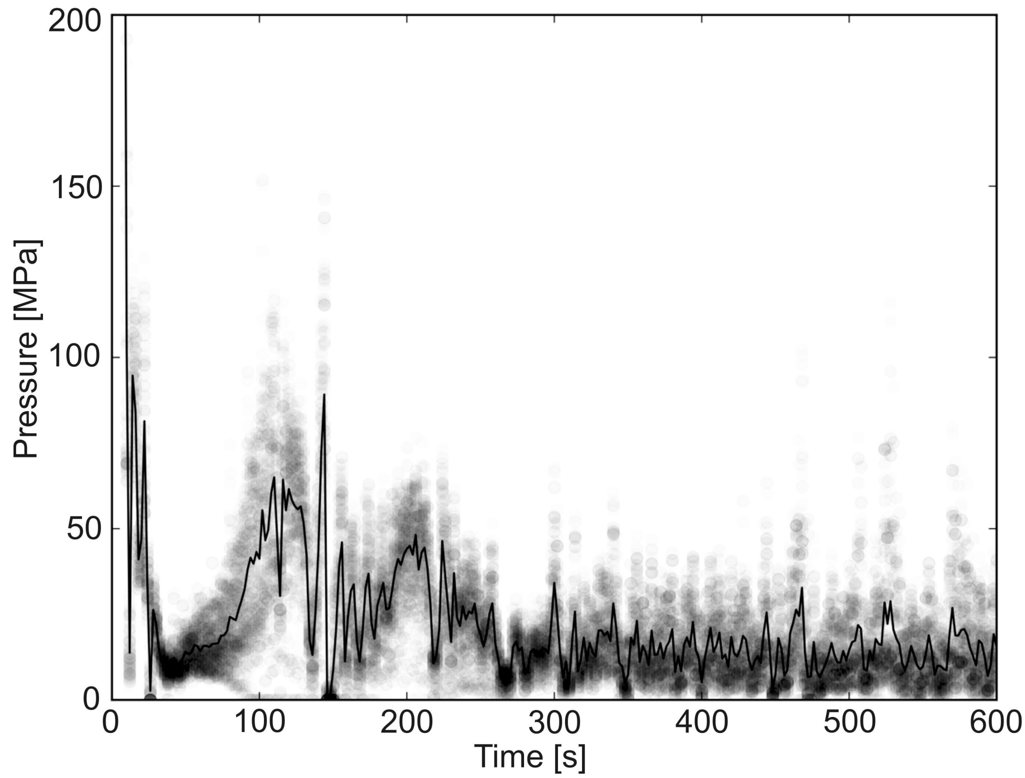
Data availability

All data generated or analysed during this study are included in this published Article. Other Expedition 364 data are available online (https://doi.org/10.14379/iodp.proc.364.2017).

38. Gulick, S. *et al*. in *Proceedings of the International Ocean Discovery Program Volume* 364 (eds Morgan, J. et al.) 1–46 (IODP, 2017).
39. Ortner, H., Reiter, F. & Acs, P. Easy handling of tectonic data: the programs Tectonics FP for Mac and Tectonics FP for Windows. *Comput. Geosci*. **28**, 1193–1200 (2002).
40. Collins, G. S. *et al*. Dynamic modeling suggests terrace zone asymmetry in the Chicxulub crater is caused by target heterogeneity. *Earth Planet. Sci. Lett*. **270**, 221–230 (2008).



Extended Data Fig. 1 | Lower-hemisphere, equal-area diagrams showing poles to pre-impact aplite, diabase and pegmatite sheet intrusions. N, north. *n*, number of dykes.



Extended Data Fig. 2 | Diagram showing pressure versus time as recorded by 100 Lagrangian tracer particles in the peak-ring rocks. (See Supplementary Video for location of tracer particles). Grey circles show the pressure of each tracer particle at time intervals of 2 s. The black solid line shows average pressure (all tracer particles). We note the elevated pressures between $T = 100$ s and $T = 250$ s during central uplift formation and collapse.

https://doi.org/10.1038/s41586-018-0594-0

Options for keeping the food system within environmental limits

Marco Springmann[1,2]*, Michael Clark[3], Daniel Mason-D'Croz[4,5], Keith Wiebe[4], Benjamin Leon Bodirsky[6], Luis Lassaletta[7], Wim de Vries[8], Sonja J. Vermeulen[9,10], Mario Herrero[5], Kimberly M. Carlson[11], Malin Jonell[12], Max Troell[12,13], Fabrice DeClerck[14,15], Line J. Gordon[12], Rami Zurayk[16], Peter Scarborough[2], Mike Rayner[2], Brent Loken[12,14], Jess Fanzo[17,18], H. Charles J. Godfray[1,19], David Tilman[20,21], Johan Rockström[6,12] & Walter Willett[22]

The food system is a major driver of climate change, changes in land use, depletion of freshwater resources, and pollution of aquatic and terrestrial ecosystems through excessive nitrogen and phosphorus inputs. Here we show that between 2010 and 2050, as a result of expected changes in population and income levels, the environmental effects of the food system could increase by 50–90% in the absence of technological changes and dedicated mitigation measures, reaching levels that are beyond the planetary boundaries that define a safe operating space for humanity. We analyse several options for reducing the environmental effects of the food system, including dietary changes towards healthier, more plant-based diets, improvements in technologies and management, and reductions in food loss and waste. We find that no single measure is enough to keep these effects within all planetary boundaries simultaneously, and that a synergistic combination of measures will be needed to sufficiently mitigate the projected increase in environmental pressures.

The global food system is a major driver of climate change[1,2], land-use change and biodiversity loss[3,4], depletion of freshwater resources[5,6], and pollution of aquatic and terrestrial ecosystems through nitrogen and phosphorus run-off from fertilizer and manure application[7–9]. It has contributed to the crossing of several of the proposed 'planetary boundaries' that attempt to define a safe operating space for humanity on a stable Earth system[10–12], in particular those concerning climate change, biosphere integrity, and biogeochemical flows related to nitrogen and phosphorous cycles. If socioeconomic changes towards Western consumption patterns continue, the environmental pressures of the food system are likely to intensify[13–16], and humanity might soon approach the planetary boundaries for global freshwater use, change in land use, and ocean acidification[11,12,17]. Beyond those boundaries, ecosystems could be at risk of being destabilized and losing the regulation functions on which populations depend[11,12].

Here we analyse the option space available for the food system to reduce its environmental impacts and stay within the planetary boundaries related to food production. We build on existing analyses that have advanced the planetary-boundary framework in terms of systemic threats to large-scale ecosystems[11,12,18–20], discussed the role of agriculture with respect to those pressures[10,21], and analysed the impacts on individual environmental domains[22,23], including selected measures to alleviate those impacts[22–24]. The planetary-boundary framework is not without criticism, particularly because of the heterogeneity of the different boundaries and their underlying scientific bases, including the difficulty of defining global ecosystem thresholds for local environmental impacts[25–27]. Despite these limitations, we consider the planetary-boundary framework to be useful for framing, in broad terms, the planetary option space that preserves the sustainability of key ecosystems. We acknowledge the ongoing debate by quantifying the planetary boundaries of the food system in terms of broad ranges that reflect methodological uncertainties (see Methods), and by reporting the environmental impacts in absolute terms (for example, emissions in tonnes of carbon dioxide equivalents), which allows for comparisons to other measures of environmental sustainability.

We advance the present state of knowledge by constructing and calibrating a global food-systems model with country-level detail that resolves the major food-related environmental impacts and includes a comprehensive treatment of measures for reducing these impacts (see Methods). The regional detail of the model accounts for different production methods and environmental impacts that are linked by imports and exports of primary, intermediate and final products. We use the food-system model and estimates of present and future food demand to quantify food-related environmental impacts at the country and crop level in 2010 and 2050 for five environmental domains and the related planetary boundaries: greenhouse-gas (GHG) emission related to climate change; cropland use related to land-system change; freshwater use of surface and groundwater; and nitrogen and phosphorus application related to biogeochemical flows.

To characterize pathways towards a food system with lower environmental impacts that stays within planetary boundaries, we connect a region-specific analysis of the food system to a detailed analysis of

[1]Oxford Martin Programme on the Future of Food, Oxford Martin School, University of Oxford, Oxford, UK. [2]Centre on Population Approaches for Non-Communicable Disease Prevention, Nuffield Department of Population Health, University of Oxford, Oxford, UK. [3]Natural Resources Science and Management, University of Minnesota, St Paul, MN, USA. [4]Environment and Production Technology Division, International Food Policy Research Institute (IFPRI), Washington, DC, USA. [5]CSIRO Agriculture and Food, Commonwealth Scientific and Industrial Research Organisation, St Lucia, Brisbane, Australia. [6]Potsdam Institute for Climate Impact Research, Potsdam, Germany. [7]CEIGRAM/Agricultural Production, Universidad Politécnica de Madrid, Madrid, Spain. [8]Environmental Systems Analysis Group, Wageningen University, Wageningen, The Netherlands. [9]WWF International, Gland, Switzerland. [10]Hoffmann Centre for Sustainable Resource Economy, Chatham House, London, UK. [11]Department of Natural Resources and Environmental Management, University of Hawai'i at Manoa, Honolulu, HI, USA. [12]Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden. [13]Beijer Institute of Ecological Economics, The Royal Swedish Academy of Sciences, Stockholm, Sweden. [14]EAT, Oslo, Norway. [15]Agricultural Biodiversity and Ecosystem Services, Bioversity International, Rome, Italy. [16]Department of Landscape Design and Ecosystem Management, Faculty of Agricultural and Food Sciences, American University of Beirut, Beirut, Lebanon. [17]Nitze School of Advanced International Studies (SAIS), Berman Institute of Bioethics, Johns Hopkins University, Baltimore, MD, USA. [18]Department of International Health of the Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [19]Department of Zoology, University of Oxford, Oxford, UK. [20]Department of Ecology, Evolution and Behavior, University of Minnesota, St Paul, MN, USA. [21]Bren School of Environmental Science and Management, University of California, Santa Barbara, CA, USA. [22]Department of Epidemiology and Department of Nutrition, Harvard T. H. Chan School of Public Health, Boston, MA, USA. *e-mail: marco.springmann@dph.ox.ac.uk

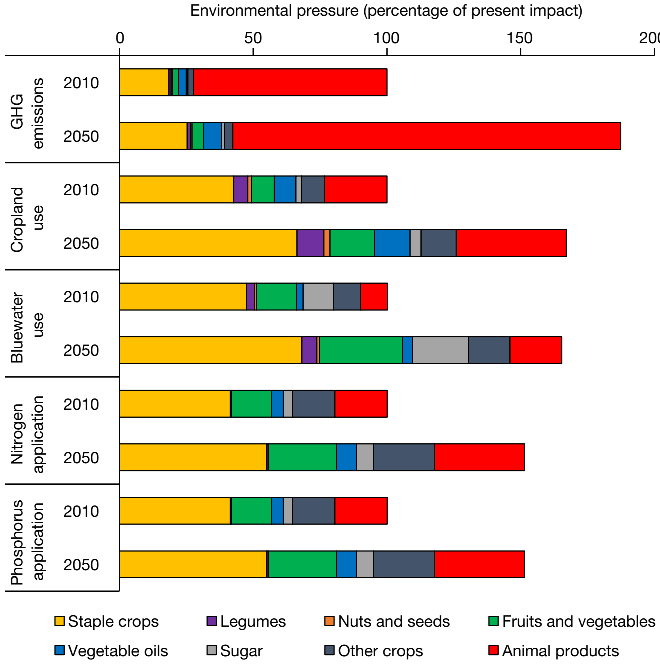


Fig. 1 | Present (2010) and projected (2050) environmental pressures on five environmental domains divided by food group. Environmental pressures are allocated to the final food product, accounting for the use and impacts of primary products in the production of vegetable oils and refined sugar, and for feed requirements in animal products. Impacts are shown as percentages of present impacts, given a baseline projection to 2050 without dedicated mitigation measures for a middle-of-the-road socioeconomic development pathway (SSP2). Absolute impacts for all socioeconomic pathways are provided in the main text and the data referred to in the 'Data availability' statement (see Methods).

measures of change, including reductions in food loss and waste, technological and management-related improvements, and dietary changes towards healthier, more plant-based diets (Extended Data Table 1). The scenarios regarding food loss and waste align with and exceed commitments made as part of the United Nations' Sustainable Development Goals[28–30]. The scenarios concerning technological change account for future improvements in agricultural yields and fertilizer application, increases in feed efficiency, and changes in management practices[31–34]. Finally, the scenarios around dietary change include changes towards dietary guidelines and more plant-based dietary patterns that are in line with present evidence on healthy eating[35–37].

In our baseline trajectory, we account for different socioeconomic pathways of population and income growth[33], and project future demand for environmental resources in the absence of technological changes and dedicated mitigation measures. Although some of the measures of change considered here can be expected to be implemented by 2050, their level of ambition is uncertain and implementation will not happen automatically. We therefore analyse each measure of change explicitly and differentiate between two degrees of implementation: medium and high ambition. Measures of medium ambition are in line with stated intentions (for example, reducing food loss and waste by half), and measures of high ambition go beyond expectations but can be considered attainable with large-scale adoption of existing best practices (for example, reducing food loss and waste by 75%).

Environmental impacts of the food system

Our analysis indicates that current and projected levels of agricultural production, in the absence of targeted mitigation measures, will greatly affect the Earth's environment. We estimate that, in 2010, the food system emitted roughly the equivalent of 5.2 billion tonnes of carbon dioxide in GHG emissions in the form of methane and nitrous oxide; the food system also occupied 12.6 million km^2 of cropland, used 1,810 km^3 of freshwater resources from surface and groundwater (bluewater), and applied 104 teragrams of nitrogen (TgN) and 18 teragrams of phosphorus (TgP) in the form of fertilizers (see Methods, 'Data availability'). Our estimates are comparable to previous estimates of food-related GHG emissions[1,38] of 4.6–5.8 billion tonnes of carbon dioxide equivalents, global cropland use[39] of 12.2–17.1 million km^2 in 2000, bluewater use[5,20] in 2000 of 1,700–2,270 km^3, and nitrogen[40] and phosphorus[40,41] application in 2010 of 104 TgN and 15.8–18.8 TgP.

Food production and consumption are projected to change between 2010 and 2050 (Extended Data Table 2) as a result of expected socioeconomic developments (Supplementary Table 1). Those developments include the growth of the global population by about a third (with a range of 23–45%, from 6.9 billion in 2010 to 8.5–10 billion in 2050) and a tripling of global income (with a range of 2.6–4.2, from US$68 trillion in 2010 to US$180–290 trillion in 2050)[33]. Because of these changes, we predict the environmental pressures of the food system to increase by 50–92% for each indicator in the absence of technological change and other mitigation measures (Fig. 1). The greatest increases along this baseline pathway are projected for GHG emissions (87%, range 80–92%), then for the demand for cropland use (67%, range 66–68%), bluewater use (65%, range 64–65%), phosphorus application (54%, range 51–55%) and nitrogen application (51%, range 50–52%).

Specific food groups vary in their environmental impacts (Fig. 1). The production of animal products generates the majority of food-related GHG emissions (72–78% of total agricultural emissions), which is due to low feed-conversion efficiencies, enteric fermentation in ruminants, and manure-related emissions[42]; the feed-related impacts of animal products also contribute to bluewater use (around 10%) and pressures on cropland, as well as nitrogen and phosphorus application (20–25% each). By comparison, staple crops have generally lower environmental footprints (impacts per kg of product) than animal products (Extended Data Table 3), in particular for GHG emissions, but they can have high total impacts because of their higher production volumes (Extended Data Table 2). According to our estimates, staple crops grown for human consumption are responsible for a third to a half (30–50%) of cropland use, bluewater use, and nitrogen and phosphorus application. The projected population growth between 2010 and 2050 contributes to a general increase in the impacts of each food group, and the projected income growth changes the relative contribution of each, with a shift towards a larger proportion of impacts from animal products (7–16% increase across environmental domains) and fruits and vegetables (2–28% increase), and a smaller proportion from staple crops (7–19% reduction).

Changes in food management, technology and diets

Reducing food loss and waste is one measure for reducing food demand and the associated environmental impacts. At present it is estimated that more than a third of all food that is produced is lost before it reaches the market, or is wasted by households[28]. For our analysis, we evaluated the impacts of reducing food loss and waste to one half—a value in line with pledges made as part of the Sustainable Development Goals[29]—and we also considered a reduction in food loss and waste by 75%, which is probably close to the maximum theoretically avoidable value[30]. We estimate that halving food loss and waste would reduce environmental pressures by 6–16% compared with the baseline projection for 2050, and that reducing food loss and waste by 75% would reduce environmental pressures by 9–24% (Fig. 2). Relatively more staple crops and fruits and vegetables are wasted than animal products[28], which explains why the impacts of changes in food loss and waste are smaller for the livestock-dominated domains, such as GHG emissions, than for the staple-crop-dominated ones, such as cropland and bluewater use and nitrogen and phosphorus application.

Technological changes increase the efficiency of production and reduce the environmental impact per unit of food produced. We analysed the most commonly considered technological advances and changes in management practices with respect to their environmental impacts (Extended Data Table 1). The measures include: increases

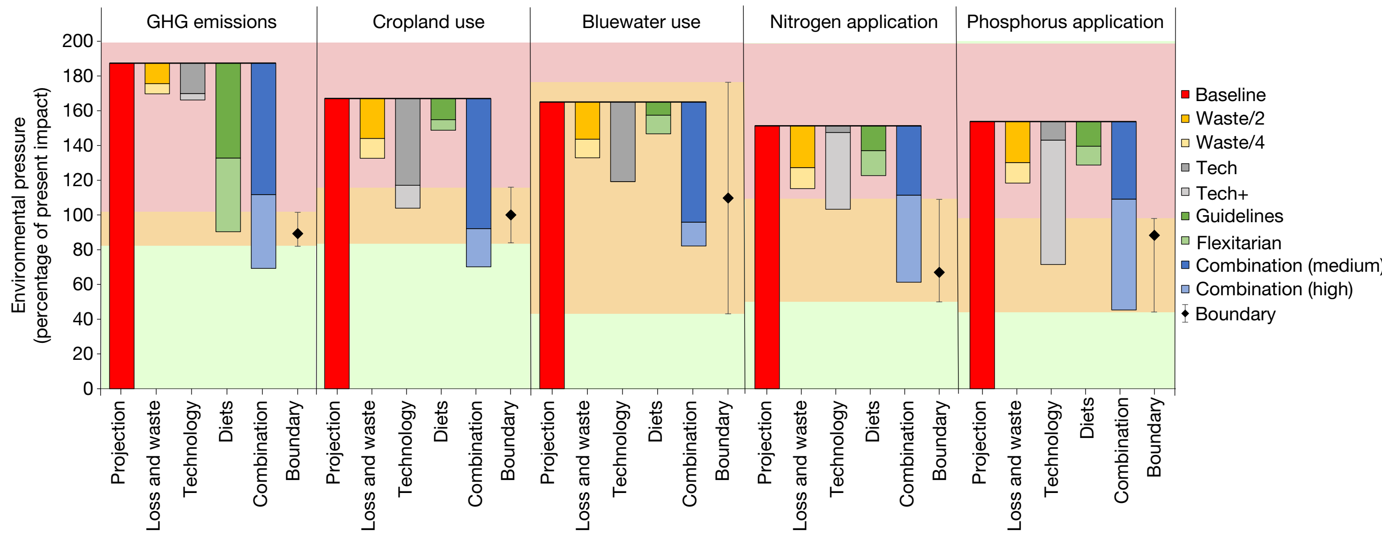


Fig. 2 | Impacts of reductions in food loss and waste, technological change, and dietary changes on global environmental pressures in 2050. These projections of environmental pressures in 2050 are baseline projections without dedicated mitigation measures for a middle-of-the-road development pathway, and are expressed as percentages of present impacts (see Fig. 1). The different measures of change and their combination are depicted as reductions from the baseline projections for the different environmental domains (for example, the 'diets' bar that ends at 90% of present impacts of GHG emissions indicates that ambitious dietary changes (flexitarian) can reduce the projected increase of GHG emissions from 187% of present impacts to 90%, which represents a reduction of 52% or 97 percentage points; and dietary changes of medium ambition (guidelines), which in the figure end at the split line of the 'diets' bar, can reduce GHG emissions from 187% of present impacts to 133%, which represents a reduction of 29% or 54 percentage points). The loss and waste scenarios include reducing food loss and waste by half (waste/2) and by 75% (waste/4). The technology scenarios include medium-ambition technological changes up to 2050 (tech) and more ambitious technological changes (tech+). The diet scenarios include diets aligned with global dietary guidelines (guidelines), and more plant-based flexitarian diets (flexitarian) that are reflective of present evidence on healthy eating. The scenario combinations include all measures of medium ambition (comb(med): waste/2, tech, guidelines) and all measures of high ambition (comb(high): waste/4, tech+, flexitarian), the latter including an optimistic socioeconomic development pathway with higher income and lower population growth. The diamonds indicate mean planetary-boundary values (boundary), each associated with uncertainty intervals highlighted by colour (light green, below the mean value; light orange, between minimum and maximum values; light red, above maximum values).

in agricultural yields, which reduce the demand for additional cropland[32,33]; rebalancing of fertilizer application between overapplying and underapplying regions[32], as well as increasing nitrogen-use efficiency[34,43] and phosphorus recycling[7], which reduce demand for additional nitrogen and phosphorus inputs; improvements in water management that increase basin efficiency, storage capacity, and better utilization of rainwater[33]; and agricultural mitigation options, including changes in irrigation, cropping and fertilization that reduce methane and nitrous oxide emissions from rice and other crops, and changes in manure management, feed conversion and feed additives that reduce enteric fermentation in livestock[31]. We estimate that implementing these measures could reduce the environmental pressures of the food system by 3–30% compared with the 2050 baseline projection in medium-ambition scenarios, and by 11–54% in high-ambition scenarios (Fig. 2). In each case, the higher-end estimates are for the staple-crop-dominated environmental indicators (cropland and bluewater use, and nitrogen and phosphorus application), for which general improvements in water management, agricultural yields, phosphorus-recycling rates and nitrogen-use efficiencies are particularly effective. The lower-end estimates are for GHG emissions, for which the contribution from livestock-related emissions is, to a large extent, an inherent characteristic of the animals and therefore cannot be reduced more substantially through existing mitigation options[31,44] (Extended Data Table 4).

Dietary changes towards healthier diets can reduce the environmental impacts of the food system when environmentally intensive foods, in particular animal products, are replaced by less intensive food types[15,16]. For our analysis, we analysed dietary changes towards diets in line with global dietary guidelines for the consumption of red meat, sugar, fruits and vegetables, and total energy intake[35,36]; as well as to more plant-based (flexitarian) diets that more comprehensively reflect the current evidence on healthy eating[37,45] by including lower amounts of red and other meats and greater amounts of fruits, vegetables, nuts and legumes (Extended Data Tables 1 and 5). We estimate that, compared with the baseline projection for 2050, dietary changes towards healthier diets could reduce GHG emissions and other environmental impacts by 29% and 5–9%, respectively, for the dietary-guidelines scenario, and by 56% and 6–22%, respectively, for the more plant-based diet scenario (Fig. 2). The changes are in line with the dietary composition of the diets and the environmental footprints of each food group (Fig. 1, Extended Data Table 1 and Supplementary Table 2). Changes in meat consumption dominate the impacts on GHG emissions, while for the other domains the environmental pressures associated with greater consumption of fruits, vegetables, nuts and legumes are more important but outweighed by the environmental benefits associated with lower consumption of meat, staple crops and sugar, and a generally lower energy intake in line with healthy body weights and recommended levels of physical activity[35] (Extended Data Table 6).

To understand how the combined implementation of some or all of the discussed measures could influence the environmental pressures of the food system, we constructed an environmental option space by combining all measures of medium ambition and all measures of high ambition. Our analysis indicates that much of the increase in environmental pressures that is expected to occur by 2050 could be mitigated if measures were combined (Fig. 2). Combining all measures of medium ambition could reduce environmental pressures by around 25–45% compared with the baseline projection for 2050, resulting in total environmental impacts that are within 15% above and below present impacts. Combining all measures of high ambition could deliver reductions of 30–60%, resulting in environmental impacts that are 20–55% less than the current ones. In line with the differentiated impacts of the different measures of change, dietary change contributes the most to the reductions in GHG emissions, and technological and management-related changes contribute the most to reductions in the other environmental impacts, while reductions in food loss and waste contribute up to a third to the overall reductions (Extended Data Fig. 1).

Diet scenario	Tech scenario	Loss and waste scenario	GHG emissions			Cropland use			Bluewater use			Nitrogen application			Phosphorus application		
			SSP2	SSP1	SSP3	SSP2	SSP1	SSP3	SSP2	SSP1	SSP3	SSP2	SSP1	SSP3	SSP2	SSP1	SSP3
Baseline	Baseline	Baseline	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4
		Waste/2	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4
		Waste/4	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4
	Tech	Baseline	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4
		Waste/2	4	4	4	3	3	3	2	2	2	4	4	4	4	4	4
		Waste/4	4	4	4	2	2	2	2	2	2	4	4	4	4	4	4
	Tech+	Baseline	4	4	4	3	3	3	3	3	3	3	3	3	2	2	2
		Waste/2	4	4	4	2	2	2	2	2	2	3	3	3	2	2	2
		Waste/4	4	4	4	1	1	1	2	2	2	3	3	3	2	2	2
Guidelines	Baseline	Baseline	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4
		Waste/2	4	4	4	4	4	4	3	3	3	4	4	4	4	4	4
		Waste/4	4	4	4	4	3	4	3	3	3	3	3	3	4	4	4
	Tech	Baseline	4	4	4	3	3	3	3	2	3	4	4	4	4	4	4
		Waste/2	4	4	4	2	2	2	2	2	2	4	3	4	4	4	4
		Waste/4	4	4	4	2	1	2	2	2	2	3	3	3	4	3	4
	Tech+	Baseline	4	4	4	2	2	2	3	2	3	3	3	3	2	2	2
		Waste/2	4	4	4	1	1	1	2	2	2	3	3	3	2	2	2
		Waste/4	4	3	4	1	1	1	2	2	2	3	3	3	2	2	2
Flexitarian	Baseline	Baseline	3	2	3	4	4	4	3	3	3	4	4	4	4	4	4
		Waste/2	1	1	2	4	4	4	3	3	3	3	3	3	4	4	4
		Waste/4	1	1	1	4	3	4	3	2	3	3	3	3	3	3	3
	Tech	Baseline	2	1	2	3	3	3	2	2	3	4	4	4	4	4	4
		Waste/2	1	1	1	2	2	2	2	2	2	3	3	3	4	4	4
		Waste/4	1	1	1	1	1	2	2	2	2	3	3	3	3	2	3
	Tech+	Baseline	1	1	2	2	2	2	2	2	3	3	3	3	2	2	2
		Waste/2	1	1	1	1	1	1	2	2	2	3	2	3	2	2	2
		Waste/4	1	1	1	1	1	1	2	2	2	2	2	2	2	1	2

Fig. 3 | Planetary option space. The figure shows combinations of dietary change, technological change (tech or tech+), changes in food loss and waste (waste/2 or waste/4), and socioeconomic development pathways (SSP1, SSP2 or SSP3). These changes are applied to baseline conditions in 2050 (baseline). The diet scenarios include diets aligned with global dietary guidelines (guidelines), and more plant-based flexitarian diets (flexitarian) that are reflective of the current evidence on healthy eating. The loss and waste scenarios include reducing food loss and waste by half (waste/2) and by 75% (waste/4). The technology scenarios include medium-ambition technological changes up to 2050 (tech) and more ambitious technological changes (tech+). The socioeconomic development pathways include a middle-of-the-road development pathway (SSP2), a more optimistic one with higher income and lower population growth (SSP1), and a more pessimistic one with lower income and higher population growth (SSP3). Colours and numbers indicate combinations that are below the lower bound of the planetary-boundary range (dark green, 1), below the mean value but above the minimum value (light green, 2), above the mean value but below the maximum (orange, 3), and above the maximum value (red, 4).

Planetary option space

What level of reduction in environmental pressures should be aimed for? We can explore this question through comparison to the associated planetary boundaries that are intended to describe a safe operating space for humanity. For our analysis, we adapted or newly quantified the food-related planetary-boundary values, including upper and lower limits (Extended Data Table 7, Extended Data Fig. 2 and Methods). According to our quantification, the planetary boundaries define a space around the present values for most environmental domains, with a mean value slightly below present values for food-related GHG emissions, at current values for cropland use, slightly above present values for bluewater use, and substantially below present values for nitrogen and phosphorus application (Fig. 2). Following the baseline trajectory of population and income change, and the related changes in food consumption and production, would lead to all mean values of the planetary boundaries being crossed. The environmental impacts of the food system would exceed the planetary boundaries for food-related GHG emissions by 110%, for cropland use by 70%, for bluewater use by 50%, for nitrogen application by 125%, and for phosphorus application by 75%.

Our analysis indicates that staying within planetary boundaries is possible with a combination of measures of high ambition for GHG emissions and nitrogen and phosphorus application, and with a combination of measures of medium ambition for cropland and bluewater use (Fig. 2). An analysis of the planetary option space details the possible combination of measures (Fig. 3). It shows that staying within the mean value of the GHG boundary requires ambitious dietary change towards more plant-based, flexitarian diets, in combination with either reductions in food loss and waste or technological improvements; staying within the mean values of the cropland and bluewater boundaries requires technological improvements in combination with reductions in food loss and waste; and staying within the mean values of the nitrogen and phosphorus boundaries requires ambitious technological improvements combined (for the nitrogen boundary) with dietary changes towards more plant-based diets, reductions in food loss and waste, and, in some combinations, a more optimistic socioeconomic development pathway that includes lower population and higher income growth than is expected at present. Combining those measures synergistically results in adoption of different measures of technological change for each environmental domain, coupled in each case to dietary changes towards more plant-based diets, reductions in food loss and waste, and an optimistic socioeconomic development pathway (Fig. 4).

Uncertainties

Our estimates are subject to several uncertainties. Some of the planetary-boundary values have a large uncertainty range, which reflects the difficulties of scaling up local environmental pressures to global levels[12,20], in particular regarding bluewater use and nitrogen and phosphorus application (see Methods). The planetary-boundary framework can therefore provide only a very broad measure of the sustainability of the food system. Our analysis indicates that using the upper bound of the planetary-boundary range increases the option space (Fig. 3) and, for example, does not require reductions in food loss and waste or a more optimistic socioeconomic development pathway; however, meeting the lower bound of the planetary-boundary range would

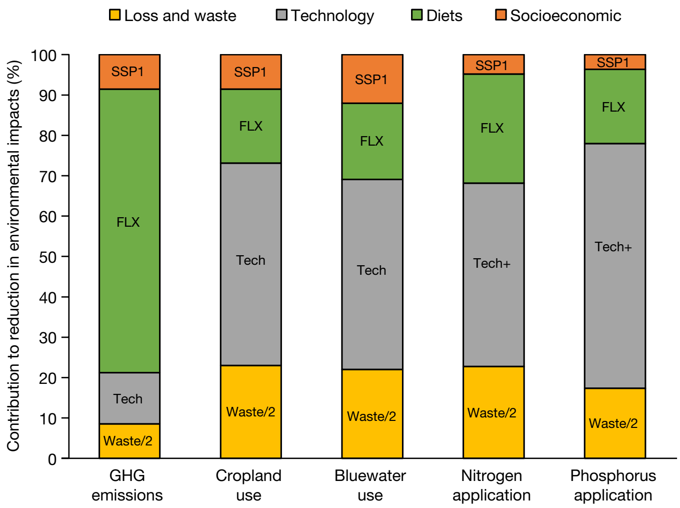


Fig. 4 | Combination and relative contributions of mitigation measures that simultaneously reduce environmental impacts below the mean values of the planetary-boundary range. The mitigation measures include different levels of technological improvements for each environmental domain (measures of high ambition (tech+) for nitrogen and phosphorus application, and measures of medium ambition (tech) for GHG emissions and for cropland and bluewater use). The other measures are not differentiated by environmental domain, and include a halving of food loss and waste (waste/2), changes towards more plant-based flexitarian diets (FLX), and optimistic socioeconomic development with higher income and lower population growth (SSP1) than expected at present. A middle-of-the-road development pathway is also feasible when combined with more ambitious reductions in food loss and waste (see Fig. 3).

not be possible for bluewater use and nitrogen application with the mitigation options considered here. Using different control variables to measure the state of planetary boundaries could also affect the option space. However, assessing the impacts of nitrogen pollution by using a measure of nitrogen surplus that accounts for all inputs and offtakes of nitrogen had little influence on the option space (Extended Data Fig. 3).

Other uncertainties are related to the set-up of our modelling framework. Although we did consider some feedback effects between the different measures of change—particularly between changes in yields and the demand for bluewater, nitrogen and phosphorus use—this was limited to the scenarios of medium ambition (see Methods). This method allowed for the differentiated adoption of ambitious technological change for domains other than cropland use without also requiring such levels for the latter. In a sensitivity analysis, we assessed the feedback effects that very high yield increases could have on nitrogen and phosphorus application[32], and found that the demand for nitrogen and phosphorus could increase across the different scenario combinations with large yield-gap closures by 8–14% and 25–32%, respectively, which would moderately reduce the planetary option space for those scenarios (Extended Data Fig. 3). In line with our focus on mitigation measures, we did not assess the impacts that climate change could have on crop yields and freshwater availability[46]. While economic responses might be able to mitigate some proportion of the biophysical impacts of climate change[47], such responses could reduce the availability and effectiveness of additional mitigation and adaptation measures, and thereby reduce the planetary option space.

Additional research would reduce the uncertainty of our scenario analysis. In our scenarios of change, we chose to focus on changes—technological, dietary, and in food loss and waste—that are considered realistic or attainable, or have been set as goals. This means that we did not include technologies or mitigation measures that have large uncertainties at present, such as soil carbon sequestration, nitrogen-fixing cereals, or landless biomass production. Some of those measures have shown some prospect in certain regions, but it is not yet clear whether they are scalable and what their relationship to existing technologies and environmental targets would be[48]. For example, land-based carbon sequestration, while reducing GHG emissions, could put additional pressures on croplands or pastures, with implications for land-use and biodiversity targets. Other areas for further research include the quantification of co-benefits of food-system change, for example, on health[15,49], biodiversity[50], and the economy[47], as well as context-specific metrics of sustainability and a greater focus on livelihood, for example in terms of food security[51].

Policy implications

Our analysis suggests that staying within the planetary boundaries of the food system requires a combination of measures: GHG emissions cannot be sufficiently mitigated without dietary changes towards more plant-based diets; cropland and bluewater use are best addressed by improvements in technologies and management that close yield gaps and increase water-use efficiency; and reducing nitrogen and phosphorus application will require a combination of measures to stay below the mean values of the planetary boundaries, including dietary change, reductions in food loss and waste, improvements in technologies and management that increase use efficiencies for nitrogen and recycling rates for phosphorus, and efforts in global socioeconomic development.

Implementation of these measures will depend on the regulatory and incentive framework in each region. In particular, practical options exist for improving technologies and management practices (Extended Data Table 1), but adoption of those options will require investment in public infrastructure, the right incentive schemes for farmers (including support mechanisms to adopt best available practices), and better regulation (for example, of water use and quality). Concrete options also exist for improving socioeconomic development in developing countries, including investments in education, particularly for women, and improving access to general and reproductive health services[52]. Meaningfully reducing food loss and waste will require measures across the entire food-supply chain[30], with possible emphasis on investments in agricultural infrastructure, technological skills, storage, transport, and distribution in developing regions; and education and awareness campaigns, food labelling, improved packaging that prolongs shelf life, and changes in legislation and business behaviour that promote closed-loop supply chains (in which waste is recycled back into the system) in developed areas. For dietary change, the available evidence suggests that providing information without additional economic or environmental changes has a limited influence on behaviour, and that integrated, multicomponent approaches that include clear policy measures might be best suited for changing diets[53,54]. Those can include a combination of media and education campaigns; labelling and consumer information; fiscal measures, such as taxation, subsidies, and other economic incentives; school and workplace approaches; local environmental changes; and direct restriction and mandates[54]. An important first step would be to align national food-based dietary guidelines with the present evidence on healthy eating and the environmental impacts of diets[55,56].

Our analysis suggests that the environmental impacts of the food system could increase markedly owing to expected changes in food consumption and production, and, in the absence of targeted measures, would exceed planetary boundaries to the extent that key ecosystem processes could become at risk of being destabilized. Synergistically combining improvements in technologies and management, reductions in food loss and waste, and dietary changes towards healthier, more plant-based diets, with particular attention to local contexts and environmental pressures, will be a key challenge in defining region-specific pathways for the sustainable development of food systems within the planetary option space. We hope that the country-specific data and suite of scenarios produced for this study (see Methods, 'Data availability') can provide a good starting point for this endeavour.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0594-0.

Received: 22 February 2018; Accepted: 30 August 2018;
Published online 10 October 2018.

1. Smith, P. et al. in *Climate Change 2014: Mitigation of Climate Change* (eds Edenhofer, O. et al.) 811–922 (Cambridge Univ. Press, 2014).
2. Vermeulen, S. J., Campbell, B. M. & Ingram, J. S. I. Climate change and food systems. *Annu. Rev. Environ. Resour.* **37**, 195–222 (2012).
3. Foley, J. A. et al. Global consequences of land use. *Science* **309**, 570–574 (2005).
4. Newbold, T. et al. Global effects of land use on local terrestrial biodiversity. *Nature* **520**, 45–50 (2015).
5. Shiklomanov, I. A. & Rodda, J. C. *World Water Resources at the Beginning of the Twenty-First Century* (Cambridge Univ. Press, Cambridge, 2004).
6. Wada, Y. et al. Global depletion of groundwater resources. *Geophys. Res. Lett.* **37**, L20402 (2010).
7. Cordell, D. & White, S. Life's bottleneck: sustaining the world's phosphorus for a food secure future. *Annu. Rev. Environ. Resour.* **39**, 161–188 (2014).
8. Diaz, R. J. & Rosenberg, R. Spreading dead zones and consequences for marine ecosystems. *Science* **321**, 926–929 (2008).
9. Robertson, G. P. & Vitousek, P. M. Nitrogen in agriculture: balancing the cost of an essential resource. *Annu. Rev. Environ. Resour.* **34**, 97–125 (2009).
10. Campbell, B. et al. Agriculture production as a major driver of the Earth system exceeding planetary boundaries. *Ecol. Soc.* **22**, 8 (2017).
11. Rockström, J. et al. A safe operating space for humanity. *Nature* **461**, 472–475 (2009).
12. Steffen, W. et al. Planetary boundaries: guiding human development on a changing planet. *Science* **347**, 1259855 (2015).
13. Davis, K. F. et al. Meeting future food demand with current agricultural resources. *Glob. Environ. Change* **39**, 125–132 (2016).
14. Jalava, M., Kummu, M., Porkka, M., Siebert, S. & Varis, O. Diet change—a solution to reduce water use? *Environ. Res. Lett.* **9**, 074016 (2014).
15. Springmann, M., Godfray, H. C. J., Rayner, M. & Scarborough, P. Analysis and valuation of the health and climate change cobenefits of dietary change. *Proc. Natl Acad. Sci. USA* **113**, 4146–4151 (2016).
16. Tilman, D. & Clark, M. Global diets link environmental sustainability and human health. *Nature* **515**, 518–522 (2014).
17. Hoekstra, A. Y. & Wiedmann, T. O. Humanity's unsustainable environmental footprint. *Science* **344**, 1114–1117 (2014).
18. Carpenter, S. R. & Bennett, E. M. Reconsideration of the planetary boundary for phosphorus. *Environ. Res. Lett.* **6**, 014009 (2011).
19. de Vries, W., Kros, J., Kroeze, C. & Seitzinger, S. P. Assessing planetary and regional nitrogen boundaries related to food security and adverse environmental impacts. *Curr. Opin. Environ. Sustain.* **5**, 392–402 (2013).
20. Gerten, D. et al. Towards a revised planetary boundary for consumptive freshwater use: role of environmental flow requirements. *Curr. Opin. Environ. Sustain.* **5**, 551–558 (2013).
21. Gordon, L. J. et al. Rewiring food systems to enhance human health and biosphere stewardship. *Environ. Res. Lett.* **12**, 100201 (2017).
22. Bodirsky, B. L. et al. Reactive nitrogen requirements to feed the world in 2050 and potential to mitigate nitrogen pollution. *Nat. Commun.* **5**, 3858 (2014).
23. Erb, K.-H. et al. Exploring the biophysical option space for feeding the world without deforestation. *Nat. Commun.* **7**, 11382 (2016).
24. Conijn, J. G., Bindraban, P. S., Schröder, J. J. & Jongschaap, R. E. E. Can our global food system meet food demand within planetary boundaries? *Agric. Ecosyst. Environ.* **251**, 244–256 (2018).
25. Lewis, S. L. We must set planetary boundaries wisely. *Nature* **485**, 417–418 (2012).
26. Montoya, J. M., Donohue, I. & Pimm, S. L. Planetary boundaries for biodiversity: implausible science, pernicious policies. *Trends Ecol. Evol.* **33**, 71–73 (2018).
27. Schlesinger, W. H. Planetary boundaries: thresholds risk prolonged degradation. *Nat. Rep. Clim. Change* **3**, 112–113 (2009).
28. Gustavsson, J., Cederberg, C., Sonesson, U., Van Otterdijk, R. & Meybeck, A. *Global Food Losses and Food Waste: Extent, Causes and Prevention* (FAO, 2011).
29. United Nations General Assembly. *Resolution Adopted by the General Assembly on 25 September 2015. 70/1 Transforming Our World: the 2030 Agenda for Sustainable Development* (United Nations, 2015).
30. Parfitt, J., Barthel, M. & Macnaughton, S. Food waste within food supply chains: quantification and potential for change to 2050. *Phil. Trans. R. Soc. B* **365**, 3065–3081 (2010).
31. Beach, R. H. et al. Global mitigation potential and costs of reducing agricultural non-CO_2 greenhouse gas emissions through 2030. *J. Integr. Environ. Sci.* **12**, 87–105 (2015).
32. Mueller, N. D. et al. Closing yield gaps through nutrient and water management. *Nature* **490**, 254–257 (2012); corrigendum 494, 390 (2013).
33. Robinson, S. et al. *The International Model for Policy Analysis of Agricultural Commodities and Trade (IMPACT)—Model Description for Version 3.* IFPRI Discussion Paper 1483 (IFPRI, 2015).
34. Sutton, M. A. et al. *Our Nutrient World: The Challenge to Produce More Food and Energy with Less Pollution* (NERC/Centre for Ecology and Hydrology, Edinburgh, UK, 2013).
35. World Health Organization. *Human Energy Requirements.* Report of a Joint FAO/WHO/UNU Expert Consultation, Rome, Italy, 17–24 October 2001 (World Health Organization, 2004).
36. World Health Organization. *Diet, Nutrition and the Prevention of Chronic Diseases.* Report of the Joint WHO/FAO Expert Consultation (World Health Organization, 2003).
37. Willett, W. C. & Stampfer, M. J. Current evidence on healthy eating. *Annu. Rev. Public Health* **34**, 77–95 (2013).
38. Tubiello, F. N. et al. The FAOSTAT database of greenhouse gas emissions from agriculture. *Environ. Res. Lett.* **8**, 015009 (2013).
39. Ramankutty, N., Evan, A. T., Monfreda, C. & Foley, J. A. Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Glob. Biogeochem. Cycles* **22**, GB1003 (2008).
40. Heffer, P., Gruère, A. & Roberts, T. *Assessment of Fertilizer Use by Crop at the Global Level 2010–2010/1.* Report A/17/134 rev (International Fertilizer Association and International Plant Nutrition Institute, 2013).
41. Zhang, J. et al. Spatiotemporal dynamics of soil phosphorus and crop uptake in global cropland during the 20th century. *Biogeosciences* **14**, 2055–2068 (2017).
42. Gerber, P. J. et al. *Tackling Climate Change through Livestock: a Global Assessment of Emissions and Mitigation Opportunities* (Food and Agriculture Organization of the United Nations, 2013).
43. Mueller, N. D. et al. Declining spatial efficiency of global cropland nitrogen allocation. *Glob. Biogeochem. Cycles* **31**, 245–257 (2017).
44. Herrero, M. et al. Greenhouse gas mitigation potentials in the livestock sector. *Nat. Clim. Change* **6**, 452–461 (2016).
45. Katz, D. L. & Meller, S. Can we say what diet is best for health? *Annu. Rev. Public Health* **35**, 83–103 (2014).
46. Rosenzweig, C. et al. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proc. Natl Acad. Sci. USA* **111**, 3268–3273 (2014).
47. Nelson, G. C. et al. Climate change effects on agriculture: economic responses to biophysical shocks. *Proc. Natl Acad. Sci. USA* **111**, 3274–3279 (2014).
48. Garnett, T. et al. *Grazed and Confused? Ruminating on Cattle, Grazing Systems, Methane, Nitrous Oxide, the Soil Carbon Sequestration Question—and What it All Means for Greenhouse Gas Emissions* (Food Climate Research Network, 2017).
49. Springmann, M. et al. Health and nutritional aspects of sustainable diet strategies and their association with environmental impacts: a global modelling analysis with country-level detail. *Lancet Planet. Health* **2**, e451–e461 (2018).
50. Tilman, D. et al. Future threats to biodiversity and pathways to their prevention. *Nature* **546**, 73–81 (2017).
51. Springmann, M. et al. Global and regional health effects of future food production under climate change: a modelling study. *Lancet* **387**, 1937–1946 (2016).
52. Abel, G. J., Barakat, B., Kc, S. & Lutz, W. Meeting the Sustainable Development Goals leads to lower world population growth. *Proc. Natl Acad. Sci. USA* **113**, 14294–14299 (2016).
53. Mozaffarian, D. Dietary and policy priorities for cardiovascular disease, diabetes, and obesity: a comprehensive review. *Circulation* **133**, 187–225 (2016).
54. Mozaffarian, D. et al. Population approaches to improve diet, physical activity, and smoking habits: a scientific statement from the American Heart Association. *Circulation* **126**, 1514–1563 (2012).
55. Ritchie, H., Reay, D. S. & Higgins, P. The impact of global dietary guidelines on climate change. *Glob. Environ. Change* **49**, 46–55 (2018).
56. Gonzales Fischer, C. & Garnett, T. *Plates, Pyramids and Planets. Developments in National Healthy and Sustainable Dietary Guidelines: a State of Play Assessment* (Univ. Oxford, 2016).

Acknowledgements This research was funded by the EAT as part of the EAT–Lancet Commission on Healthy Diets from Sustainable Food Systems, and by the Wellcome Trust, Our Planet Our Health (Livestock, Environment and People (LEAP)), award number 205212/Z/16/Z. In addition, K.W. and D.M.-D. acknowledge support from the CGIAR Research Programs on Policies, Institutions, and Markets (PIM) and on Climate Change, Agriculture and Food Security (CCAFS). K.M.C. acknowledges support from the USDA National Institute of Food and Agriculture Hatch project HAW01136-H, managed by the College of Tropical Agriculture and Human Resources. J.F. thanks Bloomberg Philanthropies and USAID for support. B.L.B. acknowledges support from the European Union's Horizon 2020 research and innovation programme under grant agreement 689150 SIM4NEXUS; the SUSTAg project; and the German Federal Minister of Education and Research (BMBF) under reference number FKZ 031B0170A. L.L. acknowledges support from MINECO, Spain, co-funded by European Commission ERDF (Ramon y Cajal fellowship, RYC-2016-20269). M.T. and M.J. thank FORMAS (grant 2016-00227). M.C. acknowledges a Balzan Award Prize and the Grand Challenge Curriculum at the University of Minnesota–Twin Cities. M.S. and H.C.J.G. acknowledge support from the Wellcome Trust, Our Planet Our Health (Livestock, Environment and People (LEAP)), award number 205212/Z/16/Z. P.S. acknowledges support from a British Heart Foundation Intermediate Basic Science Research Fellowship, FS/15/34/31656. M.R. thanks the British Heart Foundation, grant number 006/PSS/CORE/2016/OXFORD. J.R. acknowledges support from the ERC-2016-ADG 743080 (ERA).

Reviewer information *Nature* thanks K. J. Boote, G. Robertson, P. Smith and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.S. designed the study, compiled the models, conducted the analysis, interpreted the results and wrote the manuscript. K.W., D.M.-D. and M.C. contributed data and model components for the food-systems model. B.L.B., L.L. and W.d.V. contributed data and model components for

the analysis of nitrogen and phosphorus. S.J.V., M.H. and K.M.C. contributed data for the analysis of GHG emissions. M.J. and M.T. contributed data for the analysis of fish and seafood. W.W. designed the flexitarian diet and contributed to the discussion on the health aspects of dietary change. F.D. contributed to the discussion on the planetary boundary related to land use. L.J.G. and R.Z. contributed to the discussion on water use. P.S. and M.R. contributed to discussion on the health aspects of dietary change. B.L. facilitated discussions and contributed to the discussion on the planetary boundaries related to the food system. J.F. contributed to the discussion and background of the study. J.R., H.C.J.G. and D.T. contributed to discussion on the planetary boundaries related to the food system. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0594-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0594-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Food-system model. For our analysis, we constructed a food-systems model that connects food consumption and production across regions (Supplementary Information). We distinguished several steps along the food chain: primary production (including non-food uses, for example, in industry, seed banks, and as biofuels); trade in primary commodities; processing to oils, oil cakes and refined sugar; use of feed for animals; and trade in processed commodities and animals (Extended Data Table 2). We parameterized the model with data from the International Model for Policy Analysis of Agricultural Commodities and Trade (IMPACT)[33] on current and future food production, processing factors, and feed requirements for 62 agricultural commodities and 159 countries. Projections of future food consumption and production were based on statistical association with changes in income and population, and were in line with other projections[57].

To assess the environmental impacts of the food system, we paired the food-system model with a set of country-specific environmental footprints related to GHG emissions, cropland use, bluewater use, and nitrogen and phosphorus application (Extended Data Table 3; data available upon request). In line with projections of the allowable agricultural emissions budget[58], and our separate treatment of land use, we focused on the non-CO_2 emissions of agriculture, in particular methane and nitrous oxide. Data on GHG emissions were adopted from country-specific analyses of GHG emissions from crops[59] and livestock[38]. Non-CO_2 emissions of fish and seafood were calculated on the basis of feed requirements and feed-related emissions of aquaculture[60], and on projections of the ratio between wild-caught and farmed fish production[61,62]. Our baseline emissions estimate agrees well with existing ones that follow the same methodology[1,63].

Data on cropland and consumptive bluewater use were adopted from the IMPACT model[33]. To derive commodity-specific footprints, we divided use data by data on primary production, and we calculated the footprints of processed goods (vegetable oils, refined sugar) by using country-specific conversion ratios[33], and splitting co-products (oils and oil meals) by economic value to avoid double counting. We used country-specific feed requirements for terrestrial animals[33] to derive the cropland and bluewater footprints for meat and dairy, and we used global feed requirements for aquaculture[60] and projections of the ratio between wild-caught and farmed fish production[61,62] to derive the cropland and bluewater footprints for fish and seafood.

Data on fertilizer application rates of nitrogen and phosphorous were adopted from the International Fertilizer Industry Association[40]. In line with the planetary boundaries, we focus on application rates as the control variables in our main analysis. However, we note that regional environmental impacts often depend on the surplus of reactive nitrogen, a measure that accounts for all inputs and offtakes of nitrogen[64]. For a sensitivity analysis, we therefore constructed a region-specific nitrogen-budget module and linked it to the food-system model. Therein, we define the nitrogen surplus as the sum of fertilizer use, fixation by crops, manure application, human excreta and atmospheric deposition, minus nitrogen offtake by crops[22,43,65] (Supplementary Information). The results of the sensitivity analysis are reported in Extended Data Fig. 3.

Scenario analysis. We used the food-system model to estimate the environmental impacts of the food system in 2050 on GHG emissions, cropland use, bluewater use, and nitrogen and phosphorus application. To estimate the environmental impacts in the absence of dedicated mitigation measures (a scenario we term 'baseline projection'), we paired the footprints of current intensity to future projections of food demand along several socioeconomic pathways that were developed by the climate-change research community (Supplementary Table 1), including a middle-of-the-road development pathway (SSP2), a more optimistic pathway with higher income and lower population growth (SSP1), and a more pessimistic pathway with lower income and greater population growth (SSP3)[66–68]. Underlying the pathways are data and projections of the age, sex and educational structure of populations, as well as age-specific fertility, mortality and migration[67].

We then analysed the option space for reducing the environmental pressures of the food system by constructing scenarios of changes in food loss and waste, technological change and dietary change (Extended Data Table 1). For each measure, we differentiated between changes of medium and high ambition. Estimates of food loss and waste were based on percentage values reported by the UN Food and Agriculture Organization (FAO)[28]. In the standard scenario (waste/2), we assumed that food losses at the production side and food waste at the consumption side are reduced by half—a goal in line with the UN Sustainable Development Goals for 2030. In the ambitious scenarios (waste/4), we assumed reductions in food loss and waste of 75%, which is probably close to the maximum value that can be theoretically avoided[30].

The scenarios of technological change (tech and tech+) include projected efficiency gains in emissions intensities, agricultural yields, feed conversion, water use, and nitrogen and phosphorus application (Extended Data Table 4). For the scenarios describing changes in emissions intensities of foods, we incorporated the mitigation potential of bottom-up changes in management practices and technologies by using marginal abatement cost curves[31] and the value of the social cost of carbon (SCC) in 2050[69]. The mitigation options included changes in irrigation, cropping and fertilization that reduce methane and nitrous oxide emissions for rice and other crops, as well as changes in manure management, feed conversion and feed additives that reduce enteric fermentation in livestock. We used SCC values of 72 US dollars per metric ton of CO_2 equivalents (US\$/tCO_2 equivalents) associated with a rate of discounting future climate damages by 3% for the scenario of medium ambition (tech), and implemented all available mitigation options (equivalent to using a SCC of above 99 US\$/tCO_2 equivalents) for the scenario of high ambition (tech+). No marginal abatement curves were available for some crops, such as fruits, vegetables, nuts, sugar crops and oilseeds. Adopting the average mitigation potential for staple crops for these crops would increase the total mitigation potential by 1%.

Efficiency gains in agricultural yields, water management and feed conversion were based on IMPACT projections[33]. For water management, we relied on an integrated hydrological model within IMPACT that operates at the level of watersheds and accounts for management changes that increase basin efficiency, storage capacity and better utilization of rainwater[33]. For most crops, improvements in water management exceed increased water demand associated with yield improvements, except for soybeans. For agricultural yields, the gains in land-use efficiency matched estimates of yield-gap closures of about 75% between present yields and yields that are feasible in a given agricultural-climatic zone[32]. The potential efficiency gains in nitrogen and phosphorus application rates included rebalancing of fertilizer application rates between overapplying and underapplying regions in line with closing yield gaps[32]. In the ambitious technology scenario (tech+), we increased yield-gap closures to 90% on the basis of data from a previous study[32], and assumed additional improvements in nitrogen-use efficiency of 30% (in line with targets suggested by the Global Nitrogen Assessment[34]) and a recycling rate of phosphorus[7] of 50%. No further changes in efficiency were assumed for water use in the tech+ scenario. For most crops, land-use efficiencies increase in the ambitious technology scenario, except in the case of soybeans, which are assessed on a more conservative basis in a previous study[32] than by the IMPACT team.

The scenarios of dietary change include shifts towards diets that are in line with global dietary guidelines (guidelines), and towards dietary patterns that are more specialized but nutritionally balanced (flexitarian). For the former, we followed suggestions to limit the intake of red meat to less than 300 g per week[70] and the intake of added sugar to less than 5% of total energy intake (about 31 g per day)[71], to consume five portions (400 grams per day) or more of fruits and vegetables[36], and to balance energy intake (and physical activity levels) to maintain a healthy body weight[35]. Estimates of energy intake were based on the calorie needs of a moderately active population of US characteristics for height, divided into five-year age groups[72]—something that can be seen as an upper bound. Calorie needs reach a maximum of 2,500 kcal per day for ages 19–25 (averaged between men and women), but are reduced to 2,000 kcal per day for ages 66 and older. The average calorie needs differed by region according to its age composition, and averaged around 2,100 kcal per day. In a sensitivity analysis, we implemented changes in dietary composition only, without restricting energy intake. Baseline intakes of food and energy were calculated from food-availability projections of the IMPACT model by using region-specific factors of food waste and ratios of the edible portions of foods[28].

In scenarios of ambitious dietary change, we increased the stringency of the global recommendations and defined more plant-based (flexitarian) dietary patterns that reflect current evidence on healthy eating[37,46,73] (Extended Data Table 5 and Supplementary Table 2). The flexitarian diets included: at least 500 g per day of fruits and vegetables of different colours and groups (the composition of which is determined by regional preferences); at least 100 g per day of plant-based protein sources (legumes, soybeans and nuts); modest amounts of animal-based proteins, such as poultry, fish, milk and eggs; and limited amounts of red meat (one portion per week), refined sugar (less than 5% of total energy), vegetable oils that are high in saturated fat (in particular palm oil) and starchy foods with a relatively high glycaemic index. We aimed to preserve the regional character of dietary patterns by maintaining the regional composition of specific foods within broader categories, such as preferences for specific staple crops (wheat, maize, rice and so on) and fruits (temperate or tropical).

Planetary boundaries. The planetary-boundary framework attempts to define a safe operating space for humanity that is characterized by a stable Earth system[10–12]. Above planetary boundaries, it is suggested that ecosystem processes are at risk of becoming destabilized[11,12]. To contextualize the environmental impacts of the food system, we critically reviewed and adapted planetary-boundary values for GHG emissions, cropland use, bluewater use, and nitrogen and phosphorus application (Extended Data Table 7). For the climate-change boundary, we adopted an emissions budget for food-related (non-CO_2) GHG emissions that is in line with having a 66% chance of limiting global warming to below 2 °C (Representative Concentration Pathway (RCP)2.6); we derived this budget from

a model comparison of three integrated assessment models[58], normalized to the marker scenario of the associated emissions pathway[63]. The resulting budget of 4.7 $GtCO_2$ equivalents (range 4.3–5.3 $GtCO_2$ equivalents). focuses on the non-CO_2 emissions related to agriculture (methane and nitrous oxide), in line with previous assessments[58] and methodology followed by the International Panel on Climate Change. However, we note that agriculture and land use also act as source and sink for CO_2, for example through deforestation and carbon sequestration in soils[74]. How those flows should be balanced vis-à-vis the emissions from other sectors, and how additional pressures from land-based CO_2 sequestration contribute or counteract other sustainability targets and planetary boundaries, are important questions for future research.

Large uncertainties exist as to what an appropriate planetary boundary for land use should be[12]. From an analysis of forest biomes, a boundary value[12] was previously suggested in line with maintaining (not increasing pressure on) present forest cover. Such a target is in line with the strongly correlated target for biosphere integrity if nonagricultural land is placed under protection of biodiversity-compatible land use[12,75,76]. Because our modelling framework explicitly tracks cropland use, we translate the suggested target to a value of keeping current cropland use at 12.6 million km^2 (range 10.6–14.6 million km^2), given our own model calculations using the IMPACT model[33]. In future work, it will be desirable to include the role of pastures, an explicit treatment of forest cover, and further differentiation of other forms of land cover. However, a complication with switching from land use to forest cover is that the latter depends not only on agriculture, but also on wood harvesting, urbanization, and other socioeconomic variables. More than two-thirds of agricultural land is used for grazing. Converting highly productive grazing land into cropland could therefore be a conservation strategy that would relax the boundary value for cropland without affecting forest cover. However, estimates of feasible conversion ratios are still a matter of debate[23].

Two basin-level assessments of the environmental flow requirements of river systems have been used to suggest planetary boundaries for the consumption of bluewater[12,20]. We adopt the more stringent values of the more detailed standalone analysis (2,800 km^3; range 1,100–4,500 km^3)[20], which includes the other suggested values in its uncertainty range[12,77]. Because not all bluewater is used in agriculture, we scale from total consumptive bluewater use (2,550 km^3)[5] to the consumptive bluewater used in agriculture (1,810 km^3) as assessed with our hydrological model[33], which yields a boundary of 1,980 km^3 (range 780–3,190 km^3) of bluewater used in agriculture. We note that uncertainties persist about the concrete assumptions on environmental flow requirements[12,78], and about which methodology would be best suited[79].

To inform the boundary value for reactive nitrogen, a previous study[19] calculated global risk values for eutrophication on the basis of region-specific estimates of current nitrogen concentration in run-off and concentrations that would stay below ecological and toxicological thresholds of inorganic nitrogen pollution. The original boundary value for nitrogen was calculated by multiplying the global risk value by an estimate of current anthropogenic nitrogen fixation (fertilizer use plus fixation by crops)[19]. Here we apply the risk values to nitrogen application from fertilizers—in line with the focus in the planetary-boundary literature on anthropogenic disruptions of ecosystems[11,12]—and we use the nitrogen surplus (the sum of fertilizer use, fixation by crops, manure application, human excreta and atmospheric deposition, minus nitrogen offtake by crops) as a control variable in a sensitivity analysis (Extended Data Fig. 3). The resulting estimate of 52–69 TgN per year (67–90 TgN when using nitrogen surplus as a control) might be considered conservative, because the previous study[19] maintained regions that currently apply less than the critical load of nitrogen at that value, which in some cases can be much lower than needed from an environmental and food-security perspective[80]. For that reason, we adopted an upper boundary value in line with a scenario[32] that balanced nitrogen application between overapplying and underapplying regions and closed yield gaps to 75%, which yielded a final boundary value of 69 TgN (range 52–113 TgN) of nitrogen application from fertilizers (90 TgN (range 67–146 TgN) of nitrogen surplus).

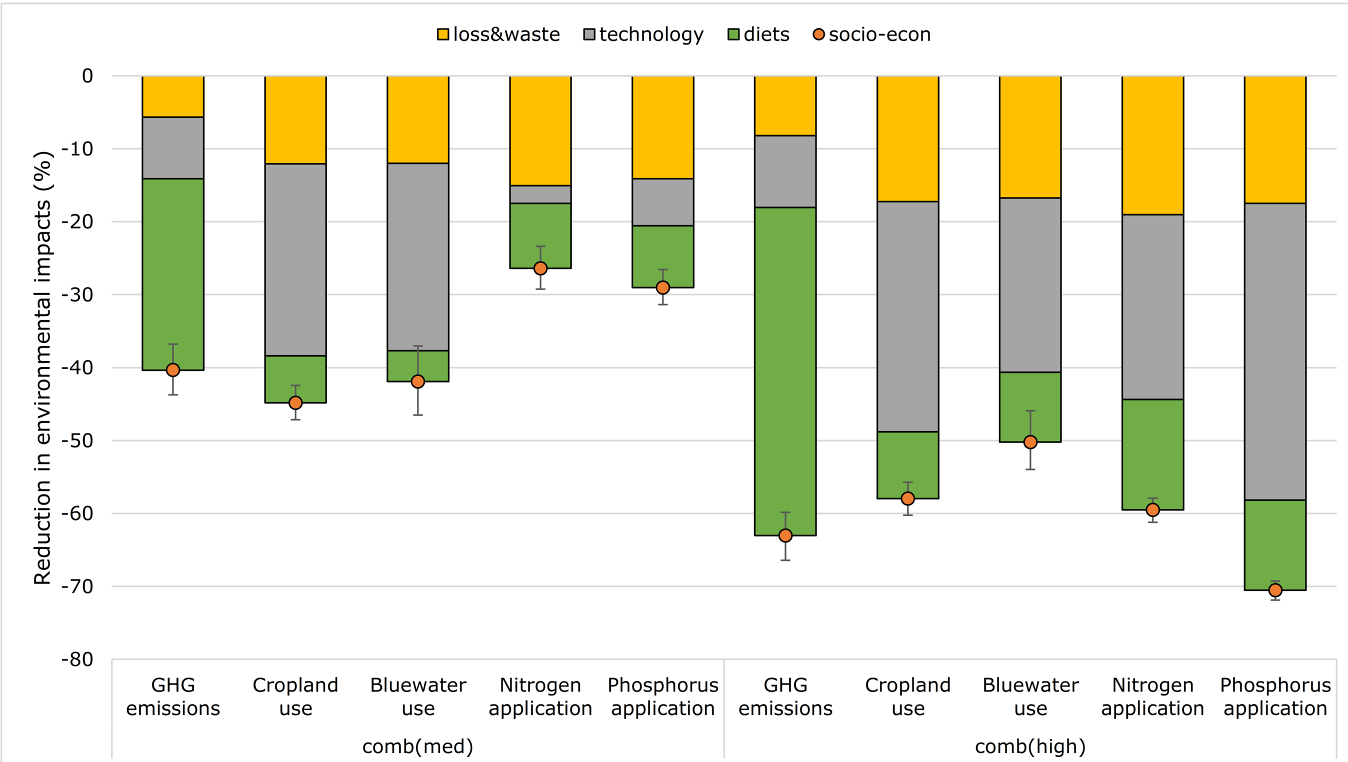
Unlike nitrogen, phosphorus can build up in the soil and is washed out as run-off during erosion[7]. Existing estimates of boundary values for phosphorus[18] have several shortcomings in that they are based on constant erosion rates and do not take into account critical sources of phosphorus, such as human waste/excreta. In the previous study[19] a global phosphorus-flow model was developed that focused on added phosphorus assuming steady-state surface pools, critical phosphorus concentrations of 50–100 mg per litre to prevent eutrophication, and flexible recycling rates (Extended Data Fig. 2 and Supplementary Information). Under no-waste recycling, the long-term phosphorus boundary amounted to 6–12 TgP per year, increasing to 8–16 TgP per year at a recycling rate of 50%. In line with our focus on scenarios of change, we adopted the latter values. As with nitrogen, there are great regional imbalances of phosphorus application[81], so we again infer an upper tolerable value from a scenario[32] that rebalanced phosphorus application between overapplying and underapplying regions and closed yield gaps to 75%. The resulting internally derived phosphorus boundary is 16 TgP (range 8–17 TgP) of phosphorus application.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

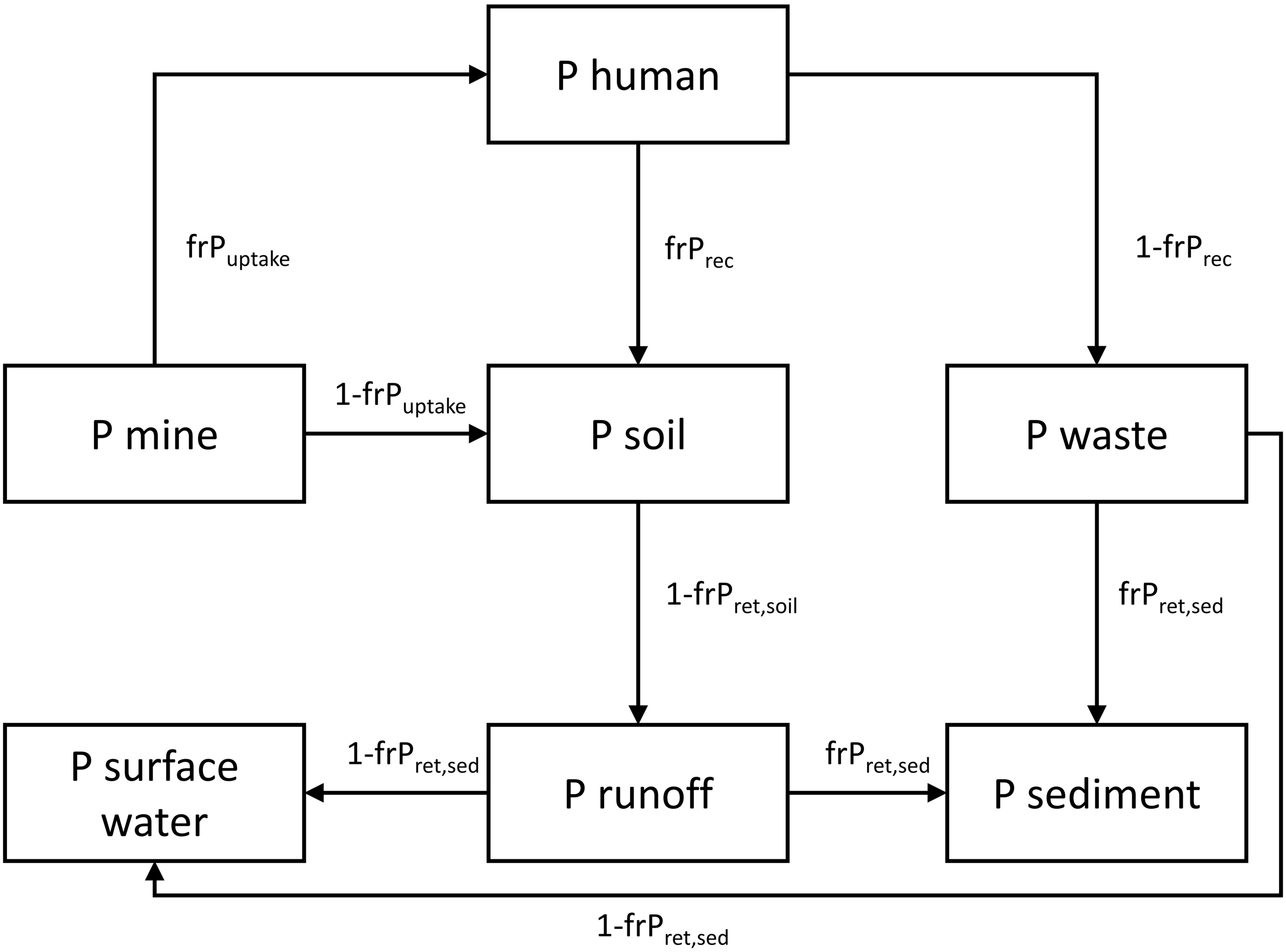
Data availability

The data that support the findings of this study are available from the Oxford University Research Archive (ORA; https://ora.ox.ac.uk) at https://ora.ox.ac.uk/objects/uuid:d9676f6b-abba-48fd-8d94-cc8c0dc546a2.

57. Alexandratos, N. & Bruinsma, J. *World Agriculture Towards 2030/2050: The 2012 Revision. ESA Working Paper No.* 12-03 (Food and Agriculture Organization of the United Nations, 2012).
58. Wollenberg, E. et al. Reducing emissions from agriculture to meet the 2 °C target. *Glob. Change Biol.* **22**, 3859–3864 (2016).
59. Carlson, K. M. et al. Greenhouse gas emissions intensity of global croplands. *Nat. Clim. Change* **7**, 63–68 (2017).
60. Troell, M. et al. Does aquaculture add resilience to the global food system? *Proc. Natl Acad. Sci. USA* **111**, 13257–13263 (2014).
61. Chan, C. Y. et al. *Fish to 2050 in the ASEAN Region* (WorldFish Center and International Food Policy Research Institute, 2017).
62. Rosegrant, M. W. et al. *Quantitative Foresight Modeling to Inform the CGIAR Research Portfolio* (International Food Policy Research Institute, 2017).
63. van Vuuren, D. P. et al. The representative concentration pathways: an overview. *Clim. Change* **109**, 5–31 (2011).
64. Fowler, D. et al. The global nitrogen cycle in the twenty-first century. *Phil. Trans. R. Soc. B* **368**, 20130165 (2013).
65. Lassaletta, L., Billen, G., Grizzetti, B., Anglade, J. & Garnier, J. 50 year trends in nitrogen use efficiency of world cropping systems: the relationship between yield and nitrogen input to cropland. *Environ. Res. Lett.* **9**, 105011 (2014).
66. Chateau, J., Dellink, R., Lanzi, E. & Magne, B. *Long-Term Economic Growth and Environmental Pressure: Reference Scenarios for Future Global Projections* (Organisation for Economic Co-operation and Development, 2012).
67. Samir, K. C. & Lutz, W. The human core of the shared socioeconomic pathways: population scenarios by age, sex, and level of education for all countries to 2100. *Glob. Environ. Change* **42**, 181–192 (2014).
68. O'Neill, B. C. et al. A new scenario framework for climate change research: the concept of shared socioeconomic pathways. *Clim. Change* **122**, 387–400 (2014).
69. Interagency Working Group on Social Cost of Greenhouse Gases. *Technical Update on the Social Cost of Carbon for Regulatory Impact Analysis—Under Executive Order 12866* (United States Government, 2013).
70. World Cancer Research Fund/American Institute for Cancer Research. *Food, Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective* (American Institute for Cancer Research, Washington DC, 2007).
71. World Health Organization. *Guideline: Sugars Intake for Adults and Children* (World Health Organization, 2015).
72. US Department of Health and Human Services & US Department of Agriculture. *Dietary Guidelines for Americans 2015-2020* 8th edn (Skyhorse Publishing, 2017).
73. Mozaffarian, D., Appel, L. J. & Van Horn, L. Components of a cardioprotective diet: new insights. *Circulation* **123**, 2870–2891 (2011).
74. Tubiello, F. N. et al. *Agriculture, Forestry and Other Land Use Emissions by Sources and Removals by Sinks: 1990–2011 Analysis.* Working Paper Series ESS 14/02 (Food and Agriculture Organization Statistical Division, 2014).
75. Dinerstein, E. et al. An ecoregion-based approach to protecting half the terrestrial realm. *Bioscience* **67**, 534–545 (2017).
76. Newbold, T. et al. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science* **353**, 288–291 (2016).
77. Rockström, J. et al. Planetary boundaries: exploring the safe operating space for humanity. *Ecol. Soc.* **14**, 32 (2009).
78. Gerten, D. et al. Response to comment on "Planetary boundaries: guiding human development on a changing planet". *Science* **348**, 1217 (2015).
79. Pastor, A. V., Ludwig, F., Biemans, H., Hoff, H. & Kabat, P. Accounting for environmental flow requirements in global water assessments. *Hydrol. Earth Syst. Sci.* **18**, 5041–5059 (2014).
80. Galloway, J. N. et al. Transformation of the nitrogen cycle: recent trends, questions, and potential solutions. *Science* **320**, 889–892 (2008).
81. MacDonald, G. K., Bennett, E. M., Potter, P. A. & Ramankutty, N. Agronomic phosphorus imbalances across the world's croplands. *Proc. Natl Acad. Sci. USA* **108**, 3086–3091 (2011).
82. Oppenheimer, M. et al. in *Climate Change 2014: Impacts, Adaptation, and Vulnerability* (eds Field, C. B. et al.) 1039–1099 (Cambridge Univ. Press, 2014).



Extended Data Fig. 1 | Reduction in environmental impacts when measures are combined. Shown are combinations of all measures of medium ambition (comb(med)) and of all measures of high ambition (comb(high)). The mitigation measures include changes in food loss and waste (loss&waste), technological change (technology) and dietary change (diets) for a middle-of-the-road development pathway. The differences to development pathways that are more optimistic (higher income and lower population growth) and more pessimistic (lower income and higher population growth) are indicated by the uncertainty range around the markers (socio-econ).



Extended Data Fig. 2 | Overview of major flows of phosphorus at the global scale. The external acceptable phosphorus (P) input is determined by the acceptable long-term accumulation of phosphorus in the soil (P soil) and sediment (P sediment) at a phosphorus concentration in surface waters (P surface water) that equals a critical threshold. The phosphorus boundary is affected by the fraction of phosphorus that is taken up by humans (P human; frP_{uptake} being the P-use efficiency, PUE, of the complete food chain, from mined phosphorus (P mine) to P intake) and the fraction of phosphorus excreted by humans (P waste) that is not recycled to land ($1 - frP_{rec}$), which becomes a point source for water pollution. This phosphorus can only be stored in sediment at a given phosphorus-retention fraction ($frP_{ret,sed}$), while the recycled phosphorus can additionally be stored in soil (at a retention fraction $frP_{ret,soil}$). The critical phosphorus input ($P_{in(crit)}$) can be calculated as the sum of critical phosphorus retention in the soil and sediment, and a critical input to surface water (oceans) that is due to run-off and leaching. The Supplementary Information contains a full derivation of phosphorus flows and quantitative estimates of critical phosphorus inputs.

Diet scenario	Tech scenario	Waste scenario	Nitrogen input (main)			Phosphorus input (main)			Nitrogen surplus			Nitrogen input (high yields)			Phosphorus input (high yields)		
			SSP2	SSP1	SSP3	SSP2	SSP1	SSP3	SSP2	SSP1	SSP3	SSP2	SSP1	SSP3	SSP2	SSP1	SSP3
BMK	BMK	BMK	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
		waste/2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
		waste/4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
	Tech	BMK	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
		waste/2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
		waste/4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
	Tech+	BMK	3	3	3	2	2	2	3	3	3	4	4	4	3	3	3
		waste/2	3	3	3	2	2	2	3	3	3	3	3	3	2	2	2
		waste/4	3	3	3	2	2	2	3	3	3	3	3	3	2	2	2
HGD	BMK	BMK	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
		waste/2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
		waste/4	3	3	3	4	4	4	4	3	4	3	3	3	4	4	4
	Tech	BMK	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
		waste/2	4	3	4	4	4	4	4	3	4	4	3	4	4	4	4
		waste/4	3	3	3	4	3	4	3	3	3	3	3	3	4	3	4
	Tech+	BMK	3	3	3	2	2	2	3	3	3	3	3	3	2	2	2
		waste/2	3	3	3	2	2	2	3	3	3	3	3	3	2	2	2
		waste/4	3	3	3	2	2	2	3	3	3	3	3	3	2	2	2
FLX	BMK	BMK	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
		waste/2	3	3	3	4	4	4	3	3	4	3	3	3	4	4	4
		waste/4	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
	Tech	BMK	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
		waste/2	3	3	3	4	4	4	3	3	3	3	3	3	4	4	4
		waste/4	3	3	3	3	2	3	3	3	3	3	3	3	3	2	3
	Tech+	BMK	3	3	3	2	2	2	3	3	3	3	3	3	2	2	2
		waste/2	3	2	3	2	2	2	3	3	3	3	3	3	2	2	2
		waste/4	2	2	2	2	1	2	3	2	3	3	2	3	2	2	2

Extended Data Fig. 3 | Planetary option space related to different control variables of nitrogen and yield-related feedback effects. The control variables include nitrogen inputs related to synthetic fertilizers as used in the main analysis, and the more comprehensive measure of nitrogen surplus that accounts for all inputs and offtakes of nitrogen. The types of feedback effects include changes in nitrogen and phosphorus application associated with closing yield gaps by 75%, as modelled in the tech scenario for cropland use (main), and changes associated with closing yield gaps by 90%, as modelled in the tech+ scenario for cropland use (high yields). Colours and numbers indicate combinations that are below the lower bound of the planetary-boundary range (dark green, 1), below the mean value but above the minimum value (light green, 2), above the mean value but below the maximum (orange, 3), and above the maximum value (red, 4).

Extended Data Table 1 | Scenarios of reductions in food loss and waste, technological change and dietary change

Scenario	Assumptions
Waste/2	Food losses and waste are reduced by half, in line with pledges made as part of the Sustainable Development Goals [29].
Waste/4	Food losses and waste are reduced by three quarters, a value likely close to the maximum value that can be theoretically avoided [30].
Tech	Closing of yield gaps between attained and attainable yields to about 75% [32,33]; Rebalancing nitrogen and phosphorus fertilizer application between over and under-applying regions [32]; improving water management, including increasing basin efficiency, storage capacity, and better utilization of rainwater [33]; and implementation of agricultural mitigation options that are economic at the projected social cost of carbon in 2050, including changes in irrigation, cropping and fertilization that reduce methane and nitrous oxide emissions for rice and other crops, as well as changes in manure management, feed conversion and feed additives that reduce enteric fermentation in livestock [31].
Tech+	Additional measures on top of TECH scenario, including additional increases in agricultural yields that close yield gaps to 90% [32]; a 30% increase in nitrogen use efficiency in line with suggested targets [34], and 50% recycling rates of phosphorus [7]; phase-out of first-generation biofuels [33]; and implementation of all available bottom-up options for mitigating food-related GHG emissions [31].
Guidelines (HGD)	Dietary shifts towards global dietary guidelines, including maximum intakes for red meat (three 100g servings per week) and sugar (5% of energy intake), minimum intakes of fruits and vegetables (five servings a day), and energy intakes in line with recommendations on healthy body weight and physical activity (2100-2200 kcal per day on average) [35,36,70,71].
Flexitarian (FLX)	Dietary shifts towards more plant-based, flexitarian dietary patterns based on recent evidence on healthy eating [37,45,73] that, in addition to the HGD requirements, include more stringent limits for red meat (one serving a week), limits for white meat (half a portion a day) and dairy (one portion a day), and greater minimum amounts of legumes, nuts, and vegetables.

HGD, guidelines; FLX, flexitarian. Data were obtained from previous studies[7,29–37,45,70,71,73].

Extended Data Table 2 | Global food production in 2010 and 2050 differentiated by food group and step along the food chain

Food item	2010								2050							
	prod	trade	intr	feed	othr	loss	waste	cons	prod	trade	intr	feed	othr	loss	waste	cons
wheat	639	114	0	109	47	34	153	295	892	256	0	138	75	51	206	423
rice	430	28	0	24	16	40	33	317	538	59	0	48	19	49	33	390
maize	797	102	0	464	96	123	32	82	1,361	304	0	878	146	156	49	133
other grains	315	45	0	172	38	30	21	55	514	146	0	268	49	55	37	106
roots	767	50	0	164	56	111	109	327	1,145	130	0	167	100	232	153	494
legumes	62	10	0	14	3	2	1	42	113	28	0	22	5	4	1	80
soybeans	225	62	189	10	8	5	0	12	357	98	282	17	22	15	1	21
nuts & seeds	37	14	8	4	4	5	0	16	54	21	9	5	8	11	1	21
vegetables	996	67	0	0	0	124	296	575	1,826	351	0	0	0	206	522	1,098
oilcrops	149	12	145	1	1	2	0	0	227	14	218	1	3	5	0	0
palmcrop	224	0	224	0	0	0	0	0	614	0	614	0	0	0	0	0
oilmeals	83	14	0	83	0	0	0	0	132	31	0	132	0	0	0	0
soybmeal	155	48	0	155	0	0	0	0	232	85	0	232	0	0	0	0
sugarcrops	1,758	0	1,758	0	0	0	0	0	3,396	0	3,396	0	0	0	0	0
fruits (temperate)	206	25	0	0	0	63	50	92	325	57	0	0	0	97	78	150
fruits (tropical)	260	37	0	0	0	26	77	156	462	94	0	0	0	54	130	277
fruits (starchy)	127	21	0	0	0	27	30	70	318	57	0	0	0	82	69	167
sugar	160	44	0	0	5	12	14	129	304	107	0	0	20	24	21	239
palm oil	45	34	0	0	1	28	0	16	124	94	0	0	2	77	1	43
vegetable oil	83	19	0	0	4	24	2	54	122	36	0	1	9	34	2	75
beef	69	9	0	0	0	1	6	63	121	20	0	0	0	1	9	111
lamb	15	2	0	0	0	0	1	13	33	5	0	0	0	0	2	30
pork	105	9	0	0	0	1	9	95	131	31	0	0	0	1	11	120
poultry	85	8	0	0	0	1	7	77	173	30	0	0	0	2	12	158
eggs	66	2	0	0	5	3	4	54	96	10	0	0	8	5	5	78
milk	627	55	0	0	0	35	35	557	1,000	156	0	0	0	67	48	885
shellfish	33	6	0	0	0	0	18	15	49	11	0	0	0	0	27	22
fish (freshwater)	41	4	0	0	0	0	22	19	81	17	0	0	0	0	43	38
fish (pelagic)	17	4	0	0	0	0	9	8	15	5	0	0	0	0	8	7
fish (demersal)	27	7	0	0	0	0	15	12	29	10	0	0	0	0	16	13

Global food production is shown in megatonnes. Steps include consumption (cons), food waste at the household level (waste), food loss at production (loss), industrial and other demand for agricultural products (othr), feed demand (feed), intermediate demand for processing into oils, oil meals and sugar (intr), traded food products (trade; globally, imports equal exports), and total production (prod = cons + waste + loss + othr + feed + intr).

Extended Data Table 3 | Environmental footprints of food commodities (per weight of product)

Food item	GHG intensity (kgCO2/kg)	Cropland use (m2/kg)	Bluewater use (m3/kg)	Nitrogen use (kgN/t)	Phosphorus use (kgP/t)
wheat	0.23	3.36	0.49	28.73	4.39
rice	1.18	3.51	1.07	36.64	5.20
maize	0.19	1.98	0.15	22.77	3.57
other grains	0.29	6.14	0.17	16.36	2.71
roots	0.07	0.69	0.04	3.63	0.71
legumes	0.23	11.02	0.95	0.00	0.00
soybeans	0.12	3.95	0.14	2.75	5.88
nuts & seeds	0.71	6.39	0.43	14.27	2.11
vegetables	0.06	0.49	0.09	9.55	1.67
fruits (temperate)	0.08	1.18	0.33	12.73	1.91
fruits (tropical)	0.09	0.94	0.32	10.27	1.58
fruits (starchy)	0.11	0.85	0.12	6.26	1.07
sugar crops	0.02	0.15	0.11	2.03	0.35
oil crops	0.46	5.45	0.31	31.33	5.61
palm crop	0.38	0.63	0.00	4.57	0.73
sugar	0.19	1.67	1.22	22.34	3.84
palm oil	1.85	3.10	0.00	22.33	3.57
vegetable oil	0.67	10.31	0.47	42.73	11.47
beef	32.49	4.21	0.22	27.29	5.36
lamb	33.02	6.24	0.49	27.51	4.94
pork	2.92	6.08	0.35	51.52	8.87
poultry	1.41	6.59	0.40	50.20	9.02
eggs	1.58	6.86	0.44	51.22	8.81
milk	1.22	1.34	0.08	6.32	1.58
shellfish	0.07	0.36	0.03	3.35	0.81
fish (freshwater)	0.30	1.51	0.10	16.78	3.62
fish (demersal)	0.02	0.12	0.01	1.20	0.29
fish (pelagic)	0.00	0.00	0.00	0.00	0.00

Footprints for animal products represent feed-related impacts, except for GHG emissions of livestock, which also have a direct component. Cropland use does not include grassland use and the use of grass inputs for ruminants. Footprints for fish and seafood represent feed-related impacts of aquaculture production weighted by total production volumes. Displayed are global averages; the regional ordering between food items can differ by region.

Extended Data Table 4 | Reductions in environmental footprints (as percentages) resulting from technological changes by food group

Food item	GHG emissions		Cropland use		Bluewater use		Nitrogen application		Phosphorus application	
	tech	tech+	tech	tech+	tech	tech+	tech	tech+	tech	tech+
wheat	-9.9	-13.8	-31.5	-37.4	-38.6	-38.6	-4.6	-33.2	-15.8	-57.9
rice	-22.4	-27.6	-25.1	-26.7	-17.2	-17.6	0.7	-29.5	-8.7	-54.3
maize	-9.7	-12.5	-32.6	-36.8	-24.6	-24.7	-10.7	-37.5	-17.8	-58.9
other grains	-10.5	-15.6	-39.6	-37.3	-27.1	-27.2	5.9	-25.9	-13.6	-56.8
roots	0.0	0.0	-32.4	-43.7	-27.5	-28.0	0.0	-30.0	0.0	-50.0
legumes	-8.8	-12.6	-38.1	-49.9	-32.3	-32.4	0.0	0.0	0.0	0.0
soybeans	-9.3	-12.4	-19.6	-2.9	10.1	9.8	0.0	-30.0	0.0	-50.0
nuts & seeds	0.0	0.0	-23.9	-35.2	-12.8	-12.8	0.0	-30.0	0.0	-50.0
vegetables	0.0	0.0	-35.4	-47.1	-39.3	-39.6	0.0	-30.0	0.0	-50.0
fruits (temperate)	0.0	0.0	-18.2	-45.5	-25.5	-25.8	0.0	-30.0	0.0	-50.0
fruits (tropical)	0.0	0.0	-35.9	-50.0	-46.1	-46.1	0.0	-30.0	0.0	-50.0
fruits (starchy)	0.0	0.0	-40.4	-57.7	-25.1	-25.1	0.0	-30.0	0.0	-50.0
sugar	0.0	0.0	-20.5	-21.2	-26.1	-26.1	0.0	-30.0	0.0	-50.0
palm oil	0.0	0.0	-22.4	-23.8	-59.1	-59.1	0.0	-30.0	0.0	-50.0
vegetable oil	-1.2	-1.5	-18.5	-43.1	-4.7	-4.9	0.0	-30.0	0.0	-50.0
beef	-9.1	-10.7	-31.4	-34.4	-25.6	-25.6	-2.2	-31.5	-13.2	-56.6
lamb	-8.6	-10.2	-35.7	-42.3	-23.9	-24.0	-1.6	-31.1	-13.2	-56.6
pork	-11.8	-15.5	-29.6	-32.6	-24.7	-25.2	-9.3	-36.5	-15.3	-57.6
poultry	-11.0	-13.6	-31.6	-34.3	-26.0	-26.1	-8.8	-36.1	-16.6	-58.3
eggs	-12.5	-15.4	-30.6	-35.3	-26.5	-26.9	-12.5	-38.8	-17.7	-58.8
milk	-9.8	-12.0	-26.3	-34.5	-16.4	-16.5	-0.2	-30.1	-5.6	-52.8
shellfish	-10.1	-12.9	-22.7	-35.3	-24.0	-24.2	-2.0	-31.4	-4.9	-52.5
fish (freshwater)	-4.7	-6.2	-22.5	-37.2	-18.9	-19.0	-3.9	-32.7	-5.6	-52.8
fish (demersal)	-5.9	-7.8	-22.5	-35.5	-22.7	-22.9	-3.8	-32.7	-5.4	-52.7
fish (pelagic)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Technological changes include changes of medium ambition (tech) and changes of high ambition (tech+). Zero entries indicate where no data were available to infer potential improvements, and for pelagic fish reflect a production method (marine fishing) that does not require feed inputs.

Extended Data Table 5 | Food-based dietary recommendations for healthy, more plant-based (flexitarian) diets

Food item	minimum level		maximum level	
	g/d	serving	g/d	serving
wheat			A total of up to 860 kcal/d for energy balance for all staple crops	3-4 (1/3 of energy)
rice				
maize				
other grains				
roots				
legumes	50	1/2		
soybeans	25	1/4		
nuts & seeds	50	2		
vegetables	300	3-4		
fruits	200	2-3		
sugar			31	5% of energy
palm oil			6.8	1
vegetable oil			80	1/3 of energy
beef			A total of 14 g/d for all red meat	1/7
lamb				
pork				
poultry			29	1/2
eggs			13	1/5
milk			250	1
shellfish	A total of 28 g/d for all fish and seafood	1/2		
fish (freshwater)				
fish (demersal)				
fish (pelagic)				

The recommendations include recommended minimum and maximum intakes expressed as weight or calories, and servings. Fish and seafood can be replaced by plant-based foods (legumes, soybeans, nuts and seeds, fruits and vegetables) in vegetarian diets. Units are g or kcal per day.

Extended Data Table 6 | Decomposition of impacts of dietary scenarios

Indicator	Diet scenario	Change in environmental impacts by food group							
		total	staples	legumes	nuts& seeds	fruits& veg	veg oils	sugar	animal products
GHG emissions (MtCO2-eq)	HGD(E=BMK)	-2,513	113	-4	-1	43	-3	-30	-2,631
	HGD	-2,850	-224	-4	-1	43	-3	-30	-2,631
	FLX	-5,063	-497	35	28	59	-7	-30	-4,651
Cropland use (1000 km2)	HGD(E=BMK)	919	1,596	0	0	450	0	-257	-870
	HGD	-1,540	-864	0	0	450	0	-257	-870
	FLX	-2,307	-2,340	1,092	407	486	716	-257	-2,415
Bluewater use (km3)	HGD(E=BMK)	201	227	0	0	244	0	-215	-56
	HGD	-136	-113	0	0	245	0	-215	-56
	FLX	-332	-394	110	39	220	48	-214	-143
Nitrogen application (GgN)	HGD(E=BMK)	3,587	10,782	0	0	2,827	1	-3,322	-6,703
	HGD	-14,784	-7,537	0	0	2,811	-4	-3,324	-6,719
	FLX	-29,723	-17,082	241	672	5,081	1,660	-3,327	-16,935
Phosphorus application (GgP)	HGD(E=BMK)	324	1,719	0	0	435	0	-570	-1,261
	HGD	-2,542	-1,136	0	0	432	-1	-571	-1,264
	FLX	-4,464	-2,670	416	118	856	607	-571	-3,212

Impacts (shown as absolute changes with respect to the baseline projection in 2050) are decomposed into changes by food group and energy intake. In the (E = BMK) scenario, only dietary composition is changed, whereas in the main scenarios, dietary composition and energy intake are changed in line with dietary guidelines and current evidence on healthy eating.

Extended Data Table 7 | Derivation of planetary-boundary values of the food system

Planetary boundary	Motivation	Method	Boundary
Climate change	Further increasing GHG emissions increase climate-related risks to ecosystems and cultures, e.g. from sea-level rise and increased occurrence of extreme weather events, such as heat waves, extreme precipitation, and coastal flooding [82].	Food-related GHG emissions in line with limiting global warming to below 2 degrees Celsius [63] with uncertainty derived from a model comparison of integrated assessment models [58].	A budget of 4.7 (4.3-5.3) $GtCO_2$-eq of food-related GHG emissions, including methane and nitrous oxide, but excluding carbon dioxide in line with IPCC methodology.
Land-system change	Further increasing the amount of agricultural land through deforestation could impact the functioning of ecosystems [3], release large amounts of carbon dioxide [1], and diminish habitat for wild species and thereby pose major threats to biodiversity [4].	Analysis of conservation levels for each forest biome in line with preserving ecosystem integrity, scaled up to a global value [12] and related to cropland use [33,39].	Not increasing pressures on forests by keeping global cropland use at 12.6 (10.6-14.6) Mkm^2. Converting productive grazing land into cropland can relax the boundary value.
Freshwater use	Further depletion and overexploitation of groundwater resources impairs natural streamflow, wetlands and related ecosystems, and can lead to land subsidence and salt-water intrusion in deltaic areas [6] and, eventually, to cascading impacts on the global hydrological cycle [77].	Basin-level assessments of the environmental flow requirements of river systems [12,20] scaled to agricultural bluewater use [5,33].	Maintaining environmental flow requirements by limiting agricultural bluewater use to 1,980 (780-3,190) km^3 or below.
Bio-geochemical flows of nitrogen and phosphorus	Agricultural runoff from overapplication of fertilizers leads to eutrophication, an increase in chemical nutrients in the water [7,9], which in turn can lead to excessive blooms of algae that deplete underwater oxygen levels resulting in so-called dead zones in coastal oceans [8].	Analysis of eutrophication risk based on nitrogen and phosphorus pollution estimates of agricultural runoff and ecological thresholds [19], with an upper value in line with re-balancing of application between over and under-applying regions [32].	Limiting nitrogen and phosphorus application from fertilizers to 69 (52-113) TgN and 16 (8-17) TgP respectively.

IPCC, Intergovernmental Panel on Climate Change. Data were obtained from previous studies[1,3–9,12,19,20,33,39,58,63,77,82].

https://doi.org/10.1038/s41586-018-0623-z

Functional genomic landscape of acute myeloid leukaemia

Jeffrey W. Tyner[1,2], Cristina E. Tognon[2,3,4], Daniel Bottomly[2,5], Beth Wilmot[2,5,6], Stephen E. Kurtz[2,3], Samantha L. Savage[1,2], Nicola Long[2,3], Anna Reister Schultz[1,2], Elie Traer[2,3], Melissa Abel[1,2], Anupriya Agarwal[2,7], Aurora Blucher[2,5], Uma Borate[2,3], Jade Bryant[1,2], Russell Burke[2,3], Amy Carlos[2,8], Richie Carpenter[2,3], Joseph Carroll[2,9], Bill H. Chang[2,10], Cody Coblentz[2,3], Amanda d'Almeida[1,2], Rachel Cook[2,3], Alexey Danilov[2,3], Kim-Hien T. Dao[2,3], Michie Degnin[2,3], Deirdre Devine[2,3], James Dibb[2,3], David K. Edwards V[1,2], Christopher A. Eide[2,3,4], Isabel English[2,3], Jason Glover[2,10], Rachel Henson[2,8], Hibery Ho[2,3], Abdusebur Jemal[2,10], Kara Johnson[2,3], Ryan Johnson[2,3], Brian Junio[2,3], Andy Kaempf[2,11], Jessica Leonard[2,3], Chenwei Lin[2,8], Selina Qiuying Liu[2,3], Pierrette Lo[2,3], Marc M. Loriaux[2,12], Samuel Luty[2,3], Tara Macey[2,3], Jason MacManiman[1,2], Jacqueline Martinez[1,2], Motomi Mori[2,11,13], Dylan Nelson[14], Ceilidh Nichols[2,3], Jill Peters[2,3], Justin Ramsdill[5,6], Angela Rofelty[1,2], Robert Schuff[5,6], Robert Searles[2,8], Erik Segerdell[2,5], Rebecca L. Smith[2,3], Stephen E. Spurgeon[2,3], Tyler Sweeney[2,3], Aashis Thapa[2,3], Corinne Visser[2,3], Jake Wagner[2,3], Kevin Watanabe-Smith[2,3], Kristen Werth[2,3], Joelle Wolf[2,10], Libbey White[2,5], Amy Yates[5,6], Haijiao Zhang[1,2], Christopher R. Cogle[15], Robert H. Collins[16], Denise C. Connolly[17,18], Michael W. Deininger[19], Leylah Drusbosky[15], Christopher S. Hourigan[20], Craig T. Jordan[21], Patricia Kropf[22], Tara L. Lin[23], Micaela E. Martinez[24], Bruno C. Medeiros[25], Rachel R. Pallapati[24], Daniel A. Pollyea[21], Ronan T. Swords[26], Justin M. Watts[26], Scott J. Weir[27,28], David L. Wiest[29], Ryan M. Winters[18], Shannon K. McWeeney[2,5,6]* & Brian J. Druker[2,3,4]*

The implementation of targeted therapies for acute myeloid leukaemia (AML) has been challenging because of the complex mutational patterns within and across patients as well as a dearth of pharmacologic agents for most mutational events. Here we report initial findings from the Beat AML programme on a cohort of 672 tumour specimens collected from 562 patients. We assessed these specimens using whole-exome sequencing, RNA sequencing and analyses of ex vivo drug sensitivity. Our data reveal mutational events that have not previously been detected in AML. We show that the response to drugs is associated with mutational status, including instances of drug sensitivity that are specific to combinatorial mutational events. Integration with RNA sequencing also revealed gene expression signatures, which predict a role for specific gene networks in the drug response. Collectively, we have generated a dataset—accessible through the Beat AML data viewer (Vizome)—that can be leveraged to address clinical, genomic, transcriptomic and functional analyses of the biology of AML.

Approximately 21,000 people are diagnosed with AML and over 10,000 AML-related deaths are reported annually in the United States[1,2]. Cytogenetic and sequencing analyses have revealed at least 11 genetic classes of AML[3] and over 20 subsets can be assigned when also considering cell differentiation states of the leukaemic blasts[4,5]. Deep sequencing of AML by The Cancer Genome Atlas (TCGA) revealed a heterogeneous disease with nearly 2,000 somatically mutated genes observed across 200 patients[6]. Many of the recurrent cytogenetic events and somatic mutations have been shown to carry prognostic importance[3,7,8]. Some of the most frequent somatic variants can also be observed in myelodysplastic syndromes and myeloproliferative neoplasms[9–11] that can transform into AML. These same mutations have also been observed in healthy individuals who have age-related clonal haematopoiesis, which is associated with significant risk for the development of myelodysplastic syndromes, myeloproliferative neoplasms and AML[12–15].

A small number of therapies targeted to mutational events have been developed for patients with AML, although the current standard of care remains largely unchanged over the past 30–40 years. The first targeted therapy for AML involved use of all-*trans* retinoic acid in combination with arsenic trioxide for patients with rearrangement of the retinoic acid receptor[16,17]. More recently, fms-related tyrosine kinase 3 (FLT3) inhibitors have been developed for *FLT3* mutational events that occur in approximately 20–30% of patients with AML[18–21].

[1]Department of Cell, Developmental & Cancer Biology, Oregon Health & Science University, Portland, OR, USA. [2]Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA. [3]Division of Hematology & Medical Oncology, Department of Medicine, Oregon Health & Science University, Portland, OR, USA. [4]Howard Hughes Medical Institute, Portland, OR, USA. [5]Division of Bioinformatics and Computational Biology, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA. [6]Oregon Clinical & Translational Research Institute, Oregon Health & Science University, Portland, OR, USA. [7]Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR, USA. [8]Integrated Genomics Laboratories, Oregon Health & Science University, Portland, OR, USA. [9]Technology Transfer & Business Development, Oregon Health & Science University, Portland, OR, USA. [10]Division of Hematology and Oncology, Department of Pediatrics, Oregon Health & Science University, Portland, OR, USA. [11]Biostatistics Shared Resource, Oregon Health & Science University, Portland, OR, USA. [12]Department of Pathology, Oregon Health & Science University, Portland, OR, USA. [13]Oregon Health & Science University-Portland State University School of Public Health, Portland, OR, USA. [14]High-Throughput Screening Services Laboratory, Oregon State University, Corvalis, OR, USA. [15]Department of Medicine, Division of Hematology and Oncology, University of Florida, Gainesville, FL, USA. [16]Department of Internal Medicine/Hematology Oncology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [17]Molecular Therapeutics Program, Fox Chase Cancer Center, Philadelphia, PA, USA. [18]Fox Chase Cancer Center Biosample Repository Facility, Philadelphia, PA, USA. [19]Division of Hematology & Hematologic Malignancies, Department of Internal Medicine, University of Utah, Salt Lake City, UT, USA. [20]National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [21]Division of Hematology, University of Colorado, Denver, CO, USA. [22]Bone Marrow Transplant Program, Fox Chase Cancer Center, Philadelphia, PA, USA. [23]Division of Hematologic Malignancies & Cellular Therapeutics, University of Kansas, Kansas City, KS, USA. [24]Clinical Research Services, University of Miami Sylvester Comprehensive Cancer Center, Miami, FL, USA. [25]Department of Medicine—Hematology, Stanford University, Stanford, CA, USA. [26]Department of Hematology, University of Miami Sylvester Comprehensive Cancer Center, Miami, FL, USA. [27]Department of Toxicology, Pharmacology and Therapeutics, University of Kansas Medical Center, Kansas City, KS, USA. [28]Department of Medicine, Division of Medical Oncology, University of Kansas Medical Center, Kansas City, KS, USA. [29]Blood Cell Development and Function Program, Fox Chase Cancer Center, Philadelphia, PA, USA. *e-mail: mcweeney@ohsu.edu; drukerb@ohsu.edu

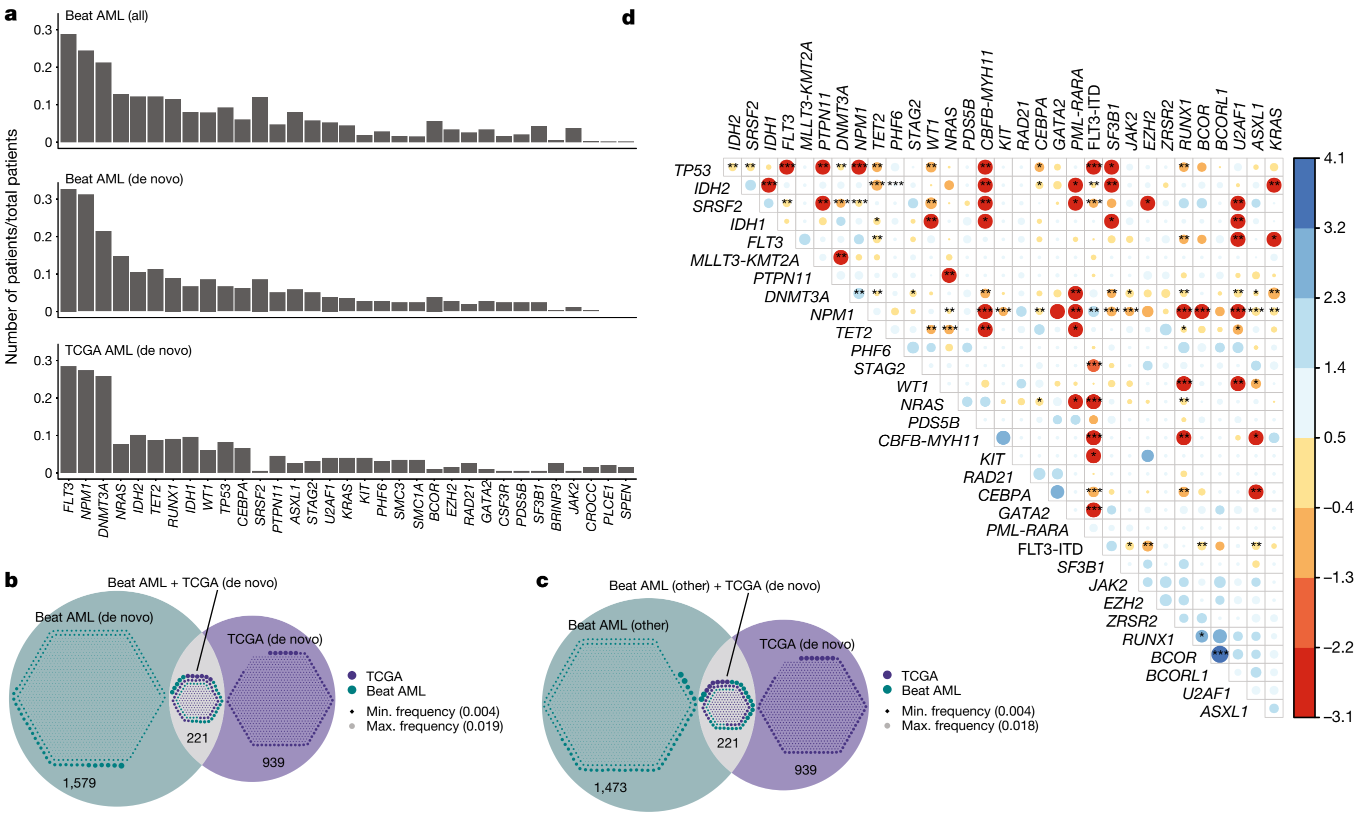


Fig. 1 | Comparative genomic landscape of AML. a, Frequency of the 33 mutational events that were cumulatively the most frequent in Beat AML ($n = 531$ patients) and TCGA ($n = 200$ patients) datasets. Top, the full Beat AML cohort; middle, only the de novo Beat AML cases; bottom, de novo cases in the TCGA. Mutations were summarized by gene as was done by TCGA, whereas the FLT3-ITD mutations were kept separate in the rest of this manuscript. **b**, Mutational events at 2% frequency or less in the de novo cases of Beat AML and TCGA were compared for overlap. Venn diagram displays the overlap with the small circles within each compartment representing a size-scaled frequency of each mutational event. **c**, Analysis as in **b** with only the non-de novo Beat AML cases versus TCGA. **d**, Co-occurrence or exclusivity of the most recurrent mutational events in the Beat AML cohort ($n = 531$ patients) were assessed using the DISCOVER[41] method. The dot plot shows the odds ratio of co-occurrence (blue) or exclusivity (red) using colour-coding and circle size as well as asterisks that indicate FDR-corrected statistical significance. $*P < 0.1$; $**P < 0.05$; $***P < 0.01$.

When FLT3 inhibitors were used as single agents, responses of only 2–6 months were obtained[22–25]. Midostaurin, a broad-spectrum FLT3 inhibitor, was recently approved for use in newly diagnosed patients with AML with *FLT3* mutations, in combination with standard of care chemotherapy[26]; however, relapse was still prevalent in this setting. Targeting of a mutant form of isocitrate dehydrogenase (NADP(+)) 1 and 2, cytosolic (IDH1 and IDH2)[27], has shown clinical benefit leading to approval of the IDH2 inhibitor, enasidenib, and the IDH1 inhibitor, ivosidenib[28,29]. Additional proposed strategies have included inhibition of epigenetic modifiers, such as enhancer of zeste 2 polycomb repressive complex 2 subunit (EZH2)[30], lysine demethylase 1A (KDM1A)[31], and DOT1-like histone lysine methyltransferase (DOT1L)[32], based on the direct mutation of these targets or synthetic lethality in the context of drug combinations (all-*trans* retinoic acid and KDM1A inhibitors) or specific genetic features (lysine methyltransferase 2A (*KMT2A*)-gene rearrangement for DOT1L inhibitors). Hypomethylating agents have been used in patients with AML, for which better responses have been reported for certain genetic subsets, such as those with mutations in *TET2*[33] or tumour protein 53 (*TP53*)[34]. Most recently, an inhibitor of the BCL2 apoptosis regulator (BCL2), venetoclax, showed an approximately 20% response rate when used as a single agent in patients with a relapse[35] and higher response rates (around 60%) were reported in combination with hypomethylating agents in newly diagnosed, elderly patients with AML[36].

Comparative genomic landscape of AML

To better understand genetic or transcriptional markers and mechanisms of drug sensitivity and resistance in AML, we developed a cohort of 672 primary specimens from 562 patients with AML and we performed extensive functional and genomic analyses on these samples. Detailed clinical annotations, including diagnostic information, clinical laboratory values, treatments, responses and outcomes were curated from electronic medical records and are reported in Supplementary Tables 1–5.

We performed exome sequencing on 622 of the specimens from the cohort representing 531 different patients. The final, high-confidence variant list (Supplementary Table 7) revealed a range of 1–80 somatic variants per patient (cohort median of 13 somatic variants) (Extended Data Fig. 1). Comparison of the top 33 most commonly mutated genes across Beat AML and TCGA[6] showed generally similar frequencies. Higher frequency of mutations in serine- and arginine-rich splicing factor 2 (SRSF2) were seen in Beat AML than in TCGA, and this difference was conserved when only the de novo cases in Beat AML and TCGA were compared (Fig. 1a). By contrast, mutational events that were seen with a frequency of less than 2% across Beat AML and TCGA were much more divergent; variants in 221 mutant genes were called in both datasets, 939 mutant genes were called only in TCGA and around 1,500 mutant genes were called only in Beat AML, irrespective of whether we compared only de novo or non-de novo cases (Fig. 1b, c). Most of these divergent mutational events were observed only in single patients; however, there were mutations in 11 genes that were called in 1% or more of patients in Beat AML, but were not observed in previous AML sequencing studies (Extended Data Fig. 1). Finally, co-occurrence and exclusivity of the most frequent variants were computed and reveal significant patterns of mutational co-segregation, suggesting biological cooperation between certain mutational events (Fig. 1d).

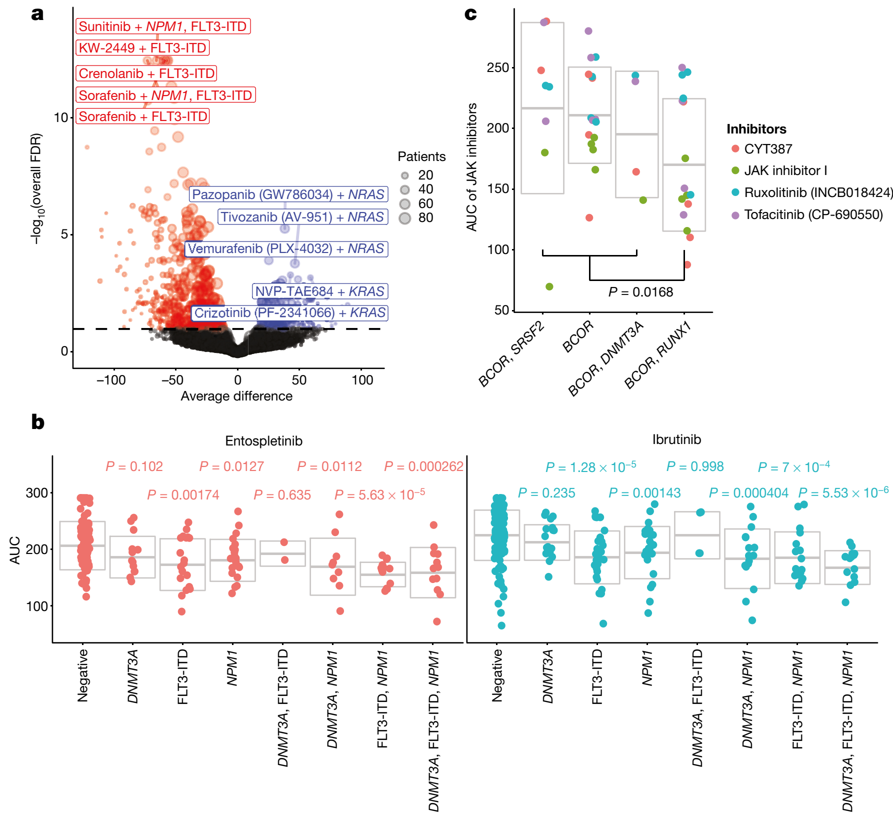


Fig. 2 | Integration of genetic events with drug sensitivity. a, Average difference in AUC drug response between mutant and wild-type cases was determined using a Student's two-sided *t*-test from a linear model fit (*x* axis). The *P* values were corrected using the Benjamini–Hochberg method over all the drugs (*y* axis). The number of samples used to correlate each mutational event with drug sensitivity is reported in the Supplementary Table 17. Expanded and interactive plots are available in our online data browser (http://www.vizome.org/ and http://vizome.org/additional_figures_BeatAML.html). **b**, AUCs for ibrutinib or entospletinib ($n = 277$ or 168 patient samples, respectively) were plotted for cases with single, double or triple mutations in *NPM1* and *DNMT3A* as well as FLT3-ITD. Data are mean ± s.d. An ANOVA was conducted using the Bonferroni approach (statistical results and sample size for all groups are reported in Supplementary Tables 18, 19). **c**, Inhibitors of JAK family kinases were assessed for activity against cases with *BCOR* mutations alone or *BCOR* mutations in combination with mutations in *SRSF2*, *RUNX1* or *DNMT3A*. The AUC values are plotted per case; data are mean ± s.d. There was a significant difference in AUC; two-sided Student's *t*-test ($t_{42} = -2.489$, $P = 0.0168$, 95% confidence interval −73.018 to −7.643) between mutations in *BCOR* and *RUNX1* ($n = 16$) versus the average of *BCOR* mutations alone ($n = 16$), mutations in *BCOR* and *SRSF2* ($n = 8$), and mutations in *BCOR* and *DNMT3A* ($n = 4$).

Drug response and gene expression

RNA sequencing was performed on 451 specimens from 411 patients. Clustering of the 2,000 most variably expressed genes across the cohort revealed gene expression signatures that were associated with many of the prominent genetic and cytogenetic disease groups (Extended Data Fig. 2). To understand the profile of sensitivity and resistance to a variety of small-molecule inhibitors, we profiled primary tumour cells from 409 specimens derived from 363 patients against a panel of 122 small-molecule inhibitors using an ex vivo drug sensitivity assay[37]. Drug sensitivity patterns were analysed with respect to clinical and genetic features of tumours (Extended Data Fig. 3). We compared the average area-under-the curve (AUC) values for each drug between samples from cases with de novo AML and cases with AML that had transformed from myelodysplastic syndromes or myeloproliferative neoplasms, through a series of single-factor analysis of variance (ANOVA) tests. Generally, transformed cases showed less sensitivity than de novo cases to most drugs. Of the 122 drugs tested, 64 were significantly (false discovery rate (FDR) < 0.1) more sensitive in the de novo samples, whereas only one drug—panobinostat (an HDAC inhibitor)—was significantly more sensitive in the transformed cases (Extended Data Fig. 4). In addition, we analysed the concordance of drug sensitivity patterns with respect to predicted target gene, gene family or pathway for each drug (drug assignments to target families can be found in Supplementary Table 11). This analysis revealed drug targets and/or drug families that were highly concordant among constituent members, as well as drug families that were quite discordant (Extended Data Fig. 5). To create a global view of overall sensitivity or resistance for each case, we generated a heat map of binary sensitive or resistant calls for each sample to each drug. We then annotated the sensitive or resistant fraction of each case against the European Leukaemia Net (ELN) 2017[5] (Extended Data Fig. 6a) and WHO (World Health Organization) 2016[4] (Extended Data Fig. 6b) classifications.

Gene signatures of drug responses

We performed a cohort analysis to assess the correlation between drug sensitivity patterns and mutational events or gene expression levels. Correlation analyses between drug sensitivity and mutational events were performed by assessing the range of sensitivity of cases with a mutation in an individual gene (as well as co-occurring mutational events) versus cases with the wild-type sequence for that same gene. Broad summaries of full cohort results are displayed in Circos and Manhattan plots (Extended Data Figs. 7a, 8). Individual associations between drugs and mutations are displayed as a Volcano plot, in which the differences in drug sensitivity between mutant and wild-type genes and the FDR-corrected significance values[38] are plotted (Fig. 2a). Some of the associations with the highest levels of statistical significance involved FLT3 internal tandem duplications (FLT3-ITD) and these showed sensitivity to FLT3 pathway inhibitors, which serves as a proof-of-principle as FLT3 inhibitors are known to be more effective against FLT3-ITD AML. However, to reveal associations between drugs and mutations that were not biased by the co-occurrence with FLT3-ITD, we also plotted the same analysis using only cases that had

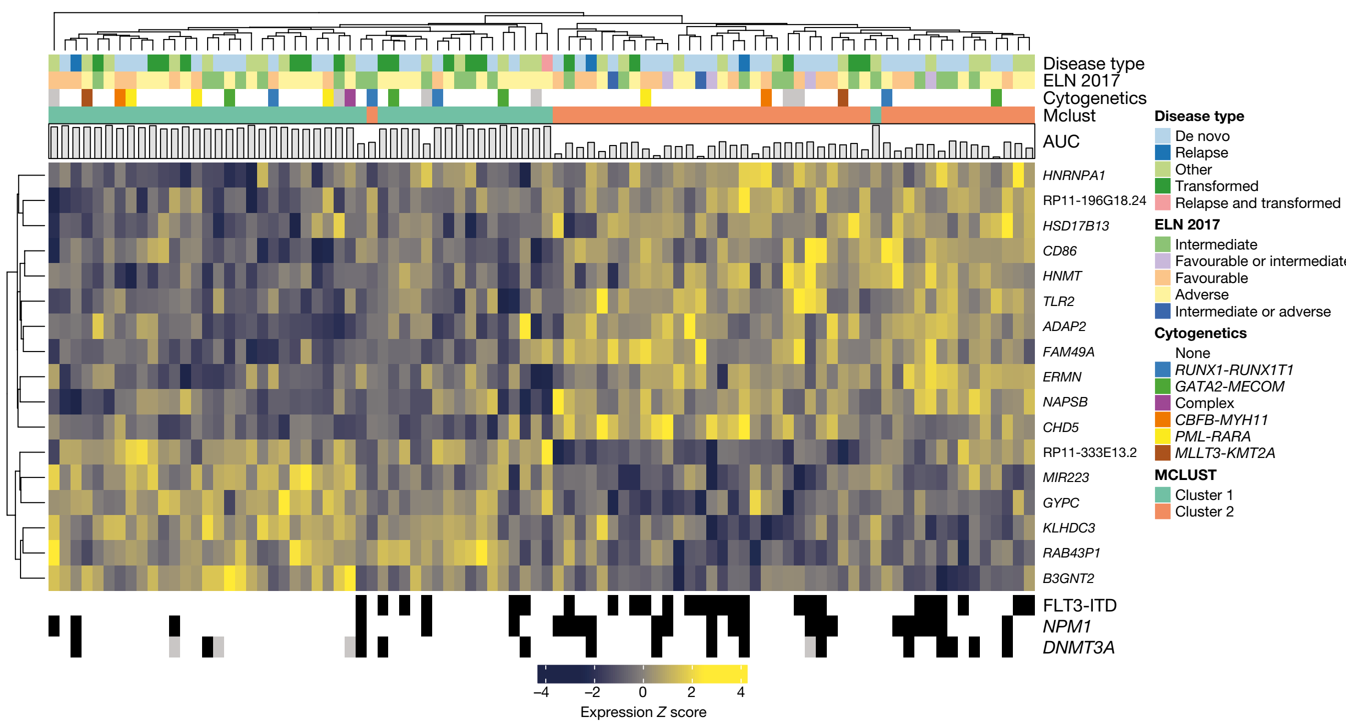


Fig. 3 | Integration of gene expression and drug sensitivity patterns. Differential gene expression signature distinguishing the 20% most ibrutinib-sensitive ($n = 46$) from the 20% most resistant ($n = 44$) specimens. Heat maps for all other drugs are available in our online data browser (http://vizome.org/additional_figures_BeatAML.html). For the number of samples used to correlate each drug with gene expression, see Supplementary Table 17. RP11-333E13.2 is a lincRNA; RP11-196G18.24 is a pseudogene.

wild-type *FLT3* (Extended Data Fig. 7b). We also created Volcano plots that were specific to each individual drug (versus all mutational events) and for each individual gene (versus all drugs tested). All Volcano plots can be found with interactive features in our online data browser (http://www.vizome.org/ and http://vizome.org/additional_figures_BeatAML.html).

Mutations in several genes, most notably *TP53* or ASXL transcriptional regulator 1 (*ASXL1*), were shown to cause a broad pattern of drug resistance. Notably, a few drugs trended to be more sensitive to cases with *TP53* mutations (such as elesclemol) or with mutations in *ASXL1* (such as panobinostat), suggesting candidates for further exploration of cases with AML that have these poor prognostic features. Mutations in the NRAS proto-oncogene, GTPase (*NRAS*) or KRAS proto-oncogene, GTPase (*KRAS*) also correlated with resistance to most drugs, although mutations in these genes did show the predicted sensitivity to MAPK inhibitors. Of particular note, there was a stronger association between mutations in *NRAS* and MAPK sensitivity than between mutations in *KRAS* and MAPK sensitivity. *IDH2* mutations conferred sensitivity to a broad spectrum of drugs, whereas mutations in *IDH1* conferred resistance to most drugs. Mutations in *RUNX1* correlated with sensitivity to PIK3C and mTOR inhibitors (such as BEZ235) and to the multi-kinase vascular endothelial growth factor receptor (VEGFR) inhibitor, cediranib. Mutation in the splicesome components U2 small nuclear RNA auxiliary factor 1 (*U2AF1*) and zinc-finger CCCH-type, RNA-binding motif and serine/arginine rich 2 (*ZRSR2*) correlated with sensitivity to several drugs. The mechanisms that underlie these latter sensitivity correlations (and many others in the dataset) merit further investigation. A significant association was seen between mutations in *FLT3*, *NPM1* and *DNMT3A* and sensitivity to the Food and Drug Administration (FDA)-approved drug, ibrutinib. Because these mutations exhibit a significant pattern of co-occurrence, we next examined every combination of single, double or triple mutant genes with respect to ibrutinib. We observed that *DNMT3A* alone or *DNMT3A* and *FLT3* double-mutant cases were not significantly different from cases with wild-type genes, whereas cases with FLT3-ITD alone or any combination with a mutation in *NPM1* (including cases in which all three genes were mutated) were significantly more sensitive than cases with wild-type genes (Fig. 2b). Ibrutinib is an inhibitor of BTK and TEC family kinases, although it can exhibit broad off-target effects when maintained in continuous culture with target cells. We noted that another kinase inhibitor with high specificity for spleen-associated tyrosine kinase (SYK)—entospletinib—showed a similarly significant pattern of sensitivity in cases with FLT3-ITD and mutations in *NPM1* (Fig. 2b), potentially pointing to an operationally important target for this disease subset. Indeed, previous studies have suggested that SYK is an interacting target of FLT3-ITD in AML[39]. Finally, we performed an additional analysis that leveraged multiple inhibitors with common targets to see whether this approach could identify additional associations. We focused on correlations between the four selective Janus kinase (JAK) inhibitors in our drug panel (momelotinib, ruxolitinib, tofacitinib and JAK inhibitor I) and mutations in the BCL6 corepressor (*BCOR*) alone or mutations in *BCOR* together with mutations in *DNMT3A*, *RUNX1* or *SRSF2*. By plotting the average difference of each JAK inhibitor between mutant and wild-type groups for these four categories and performing a one-way ANOVA of the four groups, we found that mutations in both *BCOR* and *RUNX1* correlated with increased sensitivity to all four JAK kinase inhibitors, whereas *BCOR* mutations alone or mutations in *BCOR* together with mutations in *DNMT3A* or *SRSF2* showed no difference in sensitivity to the JAK kinase inhibitors—although *BCOR* mutations alone did show sensitivity to other drugs, such as the tankyrase/WNT inhibitor XAV-939 and the multi-kinase inhibitor crizotinib (Fig. 2c). Collectively, these data suggest that dysregulation of the JAK pathway may represent a vulnerability within certain settings of specific combinations of mutations and not in others.

We also performed an integrative analysis of drug sensitivity data with respect to patterns of gene expression, comparing the 20% of samples with the lowest AUC versus the 20% with the highest AUC and assessing the most differentially expressed genes between those sample sets. This analysis revealed significant (FDR < 0.05) expression signatures for 78 testable drugs in the panel (78 out of 119; 65.5%). As an example, the 20% most and least sensitive cases to ibrutinib could be clearly distinguished by an expression signature of 17 genes

(Fig. 3 and Extended Data Fig. 9). Expression-signature heat maps for each drug can be found in our online data browser (http://vizome.org/additional_figures_BeatAML.html).

Finally, to assess the joint contributions of both mutation and system-level co-expression patterns (based on de novo network inference) to predicting drug response, multivariate modelling was performed. This integrated analysis allows us to move beyond the significant associations of single mutations (such as FLT3-ITD and mutations in *NPM1*). We performed a weighted correlation network analysis of RNA-sequencing data that identified 14 sets of genes for which the gene expression patterns showed significant clustering across the cohort (clusters contained both increased and decreased gene expression events). We performed regularized regression modelling (LASSO)[40] to understand how strongly any mutational event or any of these 14 gene expression clusters correlated with sensitivity or resistance to any of the drugs on the panel. We identified numerous, novel co-occurrences of mutations and expression clusters that were associated with drug sensitivity or resistance, with co-occurrences with ibrutinib shown as an example in Extended Data Fig. 9. For ibrutinib, these co-occurrences included a co-expression cluster of 345 genes (using a colour-labelling scheme and shown in 'turquoise') that correlated with drug sensitivity and frequently co-occurred with FLT3-ITD, which also correlated with drug sensitivity. There was significant overlap between this 'turquoise' gene expression cluster and the 17-gene signature in Fig. 3 (indicated by the ratio of observed overlap to expected overlap, or the representation factor, which was 13.6; $P < 1.734 \times 10^{-4}$). It is important to note that network analysis of this gene expression cluster highlighted enrichment of a number of immune-related pathways, which was not detected within the 17-gene signature (displayed at http://vizome.org/additional_figures_BeatAML.html). We also identified a 110-gene subnetwork (labelled in 'magenta' in Extended Data Fig. 9), which was associated with resistance to ibrutinib and was significantly associated with adverse ELN 2017 risk. To look more broadly at associations between mutations and gene expression clusters, we summarized drugs at family level to assess the frequency with which mutations and gene expression clusters were selected in iterative regression modelling (displayed at http://vizome.org/additional_figures_BeatAML.html).

Discussion

In summary, we report a large functional genomic dataset of primary tumour biopsies. We present a cohort of specimens from patients with AML for which we have performed detailed clinical annotations, genomic and transcriptomic analyses and ex vivo drug sensitivity studies, and we provide analytical approaches for data integration. Each of these datasets alone has revealed new information about the biology and potential translational approaches in AML, and the integration of these datasets has revealed new markers and mechanisms of drug sensitivity and resistance that merit further study. These data have all been made publicly available through the NIH/NCI dbGaP and Genomic Data Commons (GDC) resources, and we have developed tools to facilitate user-interfacing with the dataset (http://www.vizome.org/). We hope and expect that this public data release will stimulate further use of the data, such that novel findings can be derived and turned into new clinical approaches for treatment of AML.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0623-z.

Received: 4 April 2018; Accepted: 14 August 2018;
Published online 17 October 2018.

1. Jemal, A., Siegel, R., Xu, J. & Ward, E. Cancer statistics, 2010. *CA Cancer J. Clin*. **60**, 277–300 (2010).
2. SEER. Cancer stat facts: leukemia — acute myeloid leukemia (AML). *National Cancer Institute* https://seer.cancer.gov/statfacts/html/amyl.html (2018).
3. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med*. **374**, 2209–2221 (2016).
4. Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
5. Döhner, H. et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447 (2017).
6. The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med*. **368**, 2059–2074 (2013).
7. Byrd, J. C. et al. Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). *Blood* **100**, 4325–4336 (2002).
8. Patel, J. P. et al. Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N. Engl. J. Med*. **366**, 1079–1089 (2012).
9. Haferlach, T. et al. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia* **28**, 241–247 (2014).
10. Lundberg, P. et al. Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. *Blood* **123**, 2220–2228 (2014).
11. Deininger, M. W. N., Tyner, J. W. & Solary, E. Turning the tide in myelodysplastic/myeloproliferative neoplasms. *Nat. Rev. Cancer* **17**, 425–440 (2017).
12. Busque, L. et al. Recurrent somatic TET2 mutations in normal elderly individuals with clonal hematopoiesis. *Nat. Genet*. **44**, 1179–1181 (2012).
13. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med*. **371**, 2477–2487 (2014).
14. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med*. **371**, 2488–2498 (2014).
15. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med*. **20**, 1472–1478 (2014).
16. Huang, M. E. et al. Use of all-*trans* retinoic acid in the treatment of acute promyelocytic leukemia. *Blood* **72**, 567–572 (1988).
17. Shen, Z. X. et al. Use of arsenic trioxide (As2O3) in the treatment of acute promyelocytic leukemia (APL): II. Clinical efficacy and pharmacokinetics in relapsed patients. *Blood* **89**, 3354–3360 (1997).
18. Nakao, M. et al. Internal tandem duplication of the *FLT3* gene found in acute myeloid leukemia. *Leukemia* **10**, 1911–1918 (1996).
19. Tse, K. F., Mukherjee, G. & Small, D. Constitutive activation of FLT3 stimulates multiple intracellular signal transducers and results in transformation. *Leukemia* **14**, 1766–1776 (2000).
20. Yamamoto, Y. et al. Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies. *Blood* **97**, 2434–2439 (2001).
21. Yokota, S. et al. Internal tandem duplication of the *FLT3* gene is preferentially seen in acute myeloid leukemia and myelodysplastic syndrome among various hematological malignancies. A study on a large series of patients and cell lines. *Leukemia* **11**, 1605–1609 (1997).
22. Knapper, S. et al. A phase 2 trial of the FLT3 inhibitor lestaurtinib (CEP701) as first-line treatment for older patients with acute myeloid leukemia not considered fit for intensive chemotherapy. *Blood* **108**, 3262–3270 (2006).
23. O'Farrell, A. M. et al. An innovative phase I clinical study demonstrates inhibition of FLT3 phosphorylation by SU11248 in acute myeloid leukemia patients. *Clin. Cancer Res*. **9**, 5465–5476 (2003).
24. Smith, B. D. et al. Single-agent CEP-701, a novel FLT3 inhibitor, shows biologic and clinical activity in patients with relapsed or refractory acute myeloid leukemia. *Blood* **103**, 3669–3676 (2004).
25. DeAngelo, D. J. et al. Phase 1 clinical results with tandutinib (MLN518), a novel FLT3 antagonist, in patients with acute myelogenous leukemia or high-risk myelodysplastic syndrome: safety, pharmacokinetics, and pharmacodynamics. *Blood* **108**, 3674–3681 (2006).
26. Stone, R. M. et al. Midostaurin plus chemotherapy for acute myeloid leukemia with a *FLT3* mutation. *N. Engl. J. Med*. **377**, 454–464 (2017).
27. Mardis, E. R. et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med*. **361**, 1058–1066 (2009).
28. Wang, F. et al. Targeted inhibition of mutant IDH2 in leukemia cells induces cellular differentiation. *Science* **340**, 622–626 (2013).
29. Rohle, D. et al. An inhibitor of mutant IDH1 delays growth and promotes differentiation of glioma cells. *Science* **340**, 626–630 (2013).
30. Fiskus, W. et al. Combined epigenetic therapy with the histone methyltransferase EZH2 inhibitor 3-deazaneplanocin A and the histone deacetylase inhibitor panobinostat against human AML cells. *Blood* **114**, 2733–2743 (2009).
31. Schenk, T. et al. Inhibition of the LSD1 (*KDM1A*) demethylase reactivates the all-*trans*-retinoic acid differentiation pathway in acute myeloid leukemia. *Nat. Med*. **18**, 605–611 (2012).
32. Daigle, S. R. et al. Selective killing of mixed lineage leukemia cells by a potent small-molecule DOT1L inhibitor. *Cancer Cell* **20**, 53–65 (2011).
33. Itzykson, R. et al. Impact of *TET2* mutations on response rate to azacitidine in myelodysplastic syndromes and low blast count acute myeloid leukemias. *Leukemia* **25**, 1147–1152 (2011).
34. Welch, J. S. et al. TP53 and decitabine in acute myeloid leukemia and myelodysplastic syndromes. *N. Engl. J. Med*. **375**, 2023–2036 (2016).
35. Konopleva, M. et al. Efficacy and biological correlates of response in a phase II study of venetoclax monotherapy in patients with acute myelogenous leukemia. *Cancer Discov*. **6**, 1106–1117 (2016).
36. DiNardo, C. D. et al. Safety and preliminary efficacy of venetoclax with decitabine or azacitidine in elderly patients with previously untreated acute myeloid leukaemia: a non-randomised, open-label, phase 1b study. *Lancet Oncol*. **19**, 216–228 (2018).

37. Tyner, J. W. et al. Kinase pathway dependence in primary human leukemias determined by rapid inhibitor screening. *Cancer Res.* **73**, 285–296 (2013).
38. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
39. Puissant, A. et al. SYK is a critical regulator of FLT3 in acute myeloid leukemia. *Cancer Cell* **25**, 226–242 (2014).
40. Tibshirani, R. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1996).
41. Canisius, S., Martens, J. W. M. & Wessels, L. F. A. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol.* **17**, 261 (2016).

Acknowledgements We thank all of our patients at all sites for donating precious time and tissue. DNA and RNA quality assessments, library creation and short read sequencing assays were performed by the OHSU Massively Parallel Sequencing Shared Resource. S. Sheela, C. Lai, K. Lindblad and K. Oetjen helped with study coordination at NIH. B. Sawicki and C. Cline helped with study coordination at the University of Florida. S. Ravencroft helped with patient sample shipping and data entry and K. Schorno provided project management and support of activities at the University of Kansas Cancer Center. J. Taw helped with patient sample shipping and S. Patel helped with data entry at Stanford University. Funding for this project was provided, in part, by a Therapy Acceleration Grant to B.J.D. and J.W.T. from The Leukemia & Lymphoma Society and by support provided by the Knight Cancer Research Institute (Oregon Health & Science University, OHSU). Supported by grants from the National Cancer Institute (1U01CA217862, 1U54CA224019, 1U01CA214116, 3P30CA069533-18S5) and NIH/NCATS CTSA UL1TR002369 (S.K.M., B.W.). A.S.B. was supported by the National Library of Medicine Informatics Training Grant (T15LM007088). J.W.T. received grants from the V Foundation for Cancer Research, the Gabrielle's Angel Foundation for Cancer Research, and the National Cancer Institute (1R01CA183947). C.R.C. received a Scholar in Clinical Research Award from The Leukemia & Lymphoma Society (2400-13), was distinguished with a Pierre Chagnon Professorship in Stem Cell Biology and Blood & Marrow Transplant and a UF Research Foundation Professorship. This work was supported in part by the Intramural Research Program of the National Heart, Lung, and Blood Institute of the National Institutes of Health.

Reviewer information *Nature* thanks P. Campbell and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.W.T., C.E.T., D.B., B.W., S.E.K., S.L.S., N.L., A.R.S., E.T., S.K.M. and B.J.D. contributed equally to this work. J.W.T., C.E.T., S.K.M., A.A. and B.J.D. provided project oversight for experimental design, data management, data analysis and interpretation, methods development and development of the Vizome platform; B.W. conceived the Vizome platform, provided oversight for development of the platform and helped to provide project oversight for experimental design, workflow development, data processing, management, analysis and dissemination, and methods development; E.T. acquired patient samples, and structured, collected and analysed clinical data; S.L.S., A.R.S., M.A., I.E. and A.R. helped with processing of patient samples, ex vivo drug screening, DNA/RNA extractions and submissions of sequencing samples; N.L., R.Ca., J.D. and C.V. curated and annotated the clinical data of patients; S.E.K. provided variant confirmation and analysed data; D.B. led the workflow development for pre-processing and analysis of RNA sequencing, exome sequencing and ex vivo drug screening data, multivariate modelling and integration and helped with clinical data curation and integration, methods development and the back-end development of the Vizome platform, including the integration of Hitwalker; L.W. wrote all of the software for the Vizome platform, developed the platform as well as novel visualizations for data integration and display; J.R., R.Sc. and A.Y. provided clinical data integration from OHSU Research Data Warehouse; P.K. helped with the recruitment of patients and collection of samples for analysis; C.R.C. and R.T.S. were co-investigators for the repository protocol, and edited and provided feedback on the manuscript; L.D., C.T.J. and D.A.P. collected samples for the repository protocol and provided feedback on the manuscript; M.E.M. and R.R.P. consented patients and collected samples for the repository protocol and aided with clinical annotation; D.N. helped with the creation of drug-screening replicate plates; C.A.E., K.W.-S. and H.Z. helped with data analysis; D.K.E. analysed and curated data and developed analytical processes; A.K. and M.M. helped with the development of the ex vivo drug screening processing workflow; S.J.W. enabled, facilitated and mentored basic, translational and clinical research activities arising from data generated by Beat AML at the University of Kansas Cancer Center site, participated in the Beat AML Symposia to share research and create new research projects; A.C., R.H., C.L. and R.Se. helped with the creation of exome-sequencing and RNA-sequencing libraries and with sequencing and data processing; D.D., C.N. and J.P. helped with genomic isolation, curation of samples and entry of patient clinical annotations; E.S. helped with the whole-exome sequencing and RNA-sequencing data processing, data management and developed the data dissemination workflows; D.L.W. served as a liaison for sample acquisition; M.W.D. was a local principal investigator for the repository protocol, consented patients, collected samples for the repository protocol, aided with clinical annotation, and edited and provided feedback on the manuscript; R.M.W. served as a manager for their repository protocol, consented patients and collected samples for the repository, aided with clinical annotation, and edited and provided feedback on the manuscript; M.M.L. helped with patient sample acquisition, IRB protocol development and maintenance, clinical data structure, collection and analysis; U.M., B.H.C., R.Co., A.D., K.-H.T.D., J.L. and S.E.S. helped with the acquisition of patient samples; A.d'A., J.B., R.B., C.C., M.D., J.G., H.H., A.J., K.J., R.J., S.Q.L., S.L., J.Mac., J.Mar., R.L.S., T.S., A.T., J.Wa. and J.Wo. helped with patient sample processing and ex vivo drug screening; C.S.H. was a principal investigator for their local protocol; consented patients and collected samples for the repository protocol, aided with clinical annotation, and edited and provided feedback on the manuscript; J.M.W. was a principal investigator for their local repository protocol, and edited and provided feedback on the manuscript; B.C.M. was a principal investigator for their local repository protocol, consented patients and collected samples for the repository protocol, aided with clinical annotation, and edited and provided feedback on the manuscript; D.C.C., T.L.L. and R.H.C. were principal investigators for their local repository protocol, consented patients and collected samples for the repository protocol, aided with clinical annotation, and edited and provided feedback on the manuscript; T.M. provided regulatory oversight; B.J. and K.W. provided regulatory oversight, and helped with the curation and entry of clinical annotations of patients; A.B. provided targetome overlay; J.C. and P.L. helped with technology transfer and project development.

Competing interests J.W.T. receives research support from Agios, Aptose, Array, AstraZeneca, Constellation, Genentech, Gilead, Incyte, Janssen, Seattle Genetics, Syros and Takeda, and is a co-founder of Vivid Biosciences. M.W.D. serves on the advisory boards of and/or as a consultant for Novartis, Incyte and BMS, and receives research funding from BMS and Gilead. C.S.H. receives research funding from Merck. T.L.L. consults for Jazz Pharmaceuticals and receives research funding from Tolero, Gilead, Precient, Ono, Bio-Path, Mateon, Genentech/Roche, Trovagene, Abbvie, Pfizer, Celgene, Imago, Astellas, Karyopharm, Seattle Genetics and Incyte. D.A.P. receives research funding from Pfizer and Agios and served on the advisory boards of Pfizer, Celyad, Agios, Celgene, AbbVie, Argenx, Takeda and Servier. B.J.D. serves on the advisory boards of Gilead, Aptose, and Blueprint Medicines and is a principal investigator or coinvestigator on Novartis and BMS clinical trials. The Oregon Health & Science University (on behalf of B.J.D.) has contracts with these companies to pay for patient costs, nurse and data manager salaries and institutional overhead. B.J.D. does not derive salary, nor does his laboratory receive funds from these contracts. The authors certify that all compounds tested in this study were chosen without input from any of the industry partners.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0623-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0623-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.K.M. or B.J.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Patient samples. All patients gave informed consent to participate in this study, which had the approval and guidance of the Institutional Review Boards (IRB) at Oregon Health & Science University (OHSU), University of Utah, University of Texas Medical Center (UT Southwestern), Stanford University, University of Miami, University of Colorado, University of Florida, National Institutes of Health (NIH), Fox Chase Cancer Center and University of Kansas (KUMC). Samples were sent to the coordinating centre (OHSU; IRB 9570; 4422; NCT01728402), where they were coded and processed. Specific names of centres associated with each specimen were coded and centres providing less than five samples were aggregated together and given one centre identifier. Mononuclear cells were isolated by Ficoll gradient centrifugation from freshly obtained bone marrow aspirates or peripheral blood draws. Cell pellets were snap-frozen in liquid nitrogen for subsequent DNA isolation (Qiagen, DNeasy Blood & Tissue Kit), freshly pelleted cells were lysed immediately in guanidinium thiocyanate (GTC) lysate for subsequent RNA isolation (Qiagen, RNeasy Mini Kit), and freshly isolated mononuclear cells were plated into an ex vivo drug sensitivity assays within 24 h (described in 'Ex vivo functional drug screens'). Skin punch biopsies were collected at the site of Jamshidi needle insertion for subsequent bone marrow biopsies and genomic DNA was isolated for use as matched normal controls for exome sequencing (Qiagen, DNeasy Blood & Tissue Kit). Clinical, prognostic, genetic, cytogenetic and pathologic laboratory values as well as treatment and outcome data were manually curated from the electronic medical records of the patient. Patients were assigned a specific diagnosis based on the prioritization of genetic and clinical factors as determined by WHO guidelines. To prevent re-identification, any patient over the age of 90 was placed into a >90 aggregated age bracket. Genetic characterization of the leukaemia samples included results of a clinical deep-sequencing panel of genes commonly mutated in haematologic malignancies (Sequenome and GeneTrails (OHSU); Foundation Medicine (UTSW); Genoptix; and Illumina).

Whole-exome sequencing and custom-capture validation sequencing. For whole-exome sequencing, Illumina Nextera RapidCapture Exome capture probes and protocol were used, which provided coverage of 37 Mb of genomic DNA-coding regions. In brief, following initial quality control on a TapeStation (Agilent), 50 ng of intact genomic DNA was fragmented and tagged (tagmentation) in a single step. Following clean-up, the tagmented DNA was amplified by 10 cycles of PCR, which added the indexed adaptors for clustering and sequencing. Libraries were hybridized to capture pools in 12 sample sets with two rounds of hybridization performed to increase specificity. Libraries recovered with streptavidin magnetic beads were amplified by 10 cycles of PCR, unincorporated reagents were removed with AMPure beads (Agencourt), and validated on the TapeStation. Quantification of capture pools was done using real-time PCR (Kapa). Libraries were denatured, flow cells set up using the cBot (Illumina), and run on a HiSeq 2500 using paired-end 100-cycle protocols. For the AML tumour samples, 5 or 6 lanes were run per capture group. For the matched skin biopsy samples, 3 lanes (or equivalent) were run per capture group. The instrument and chemistry for all capture groups are provided in Supplementary Tables 12, 13.

For validation of sequencing results, an 11.9-Mb custom-capture library was developed that provided coverage of all variants previously reported in AML as well as all new variants detected from exome sequencing in this study (genes, variants and capture regions for this custom library can be found in Supplementary Table 14). This capture library was then applied to sequence 96 specimens that had previously been subjected to whole-exome sequencing for validation of variant calls.

Whole-exome sequencing data processing. We developed customized analytical pipelines that combined published algorithms with novel filtering, curation and quality-control steps. Detailed depictions of analytical workflows can be found in Supplementary Table 6 and on our online browser (http://vizome.org/additional_figures_BeatAML.html). Initial data processing and alignments were performed with commonly used analytical tools. For each flowcell and each sample, the FASTQ files were aggregated into single files for read 1 and 2. During this process, these reads were trimmed by 3 on the 5′ end and 5 on the 3′ end. BWA MEM version 0.7.10-r789[42] was used to align the read pairs for each sample-lane FASTQ file. As part of this process, the flowcell and lane information was kept as part of the read group of the resulting SAM file. The Genome Analysis Toolkit (version 3.3) and the bundled Picard (version 1.120.1579) were used[43] for alignment post-processing. The files contained within the Broad's bundle 2.8 were used including their version of the build 37 human genome (these files were downloaded from ftp://ftp.broadinstitute.org/bundle/2.8/b37/). The following steps were performed per sample-lane SAM file generated for each CaptureGroup: (1) the SAM files were sorted and converted to BAM via SortSam; (2) MarkDuplicates was run, marking both lane level standard and optical duplicates; (3) the reads were realigned around insertions and deletions (indels) from the reads using RealignerTargetCreator/IndelRealigner; (4) base quality score recalibration was performed.

The resulting BAM files were then aggregated by sample and an additional round of MarkDuplicates was performed out at the sample level. Quality-control reports were generated using the ReportingTools[44] and qrqc[45] Bioconductor R packages along with sequencing core and alignment output files. Each AML sample BAM was paired with its skin biopsy pair and an additional round of indel realignment was carried out to ensure consistency of genotypes between the two samples. If an AML sample did not have a pair, the indel realignment was instead done at the sample level.

Whole-exome sequencing variant detection. For genotyping, each AML–skin biopsy pair was realigned at the sample level and then genotyped for single-nucleotide variations using Mutect version 1.1.7[46] and Varscan2 version 2.4.1[47]. Indels were produced using Varscan2. Each variant call format (VCF) file was annotated using the Variant Effect Predictor version 83 against GRCh37[48]. The resulting VCF files were filtered to include only those annotated to a gene and were converted to mutation annotation format (MAF) format using the vcf2maf version 1.6.6 tool[49].

Mutect version 1.1.7[46] was run using default parameters, except that no limit was placed on the number or frequency of the alternative allele frequency in the normal condition to help to address normal contamination.

Varscan2 version 2.4.1[47] was run in somatic mode with the recommended filtering scheme[50], except as shown in Supplementary Table 23.

Indels and single-nucleotide variants (SNVs) were produced for the tumour-only samples again using Mutect without a specified normal for consistency and VarScan2 in mpileup2indel or mpileup2snp mode, respectively. These variants were assigned to their most deleterious effect on Ensembl transcripts using Ensembl VEP version 83 on GRCh37. This assignment was done using the same VEP parameters as the vcf2maf (version 1.5.0) program.

The TCGA AML variants[6] in MAF form were downloaded from the GDC archive site: https://portal.gdc.cancer.gov/legacy-archive/files/c410d927-d49c-4d4f-8356-601bee563ebe. The MAF files were converted to VCF files using the vcf2maf suite[51]. The resulting VCF files were lifted over from NCBI36 to GRCh37 of the human genome using CrossMap[52]. Only those variants that successfully lifted over were kept.

Variants from supplementary table 2 from a previously published study[14] were extracted and further processed, removing variants that were ambiguous in terms of external sources and could not be found in their whole-exome sequencing variants. The unique set of Beat AML variants was annotated relative to RefSeq transcripts using Ensembl VEP similar to above and all consequences were kept. This set of variants and consequences was searched against the set of processed variants from the previously published study[14].

Using the runs from MuTect and VarScan2, these data were next filtered to keep only the protein-influencing SNVs and indels from Mutect and VarScan2 and filtered, requiring that the variants had at least 5 reads and either not be seen in the Exome Aggregation Consortium (ExAC)[53] dataset or be seen at a frequency <0.01. These data present several additional challenges. First, somatic calls cannot be obtained directly from the tumour-only samples. Second, there is always a possibility of tumour contamination of the skin samples for those samples that were paired. To address these issues and maximize comparability, we used an iterative approach. The following was done separately for the two genotypers. (1) An initial set of higher-confidence somatic mutations were retrieved from the paired tumour–normal samples requiring tumour variant allele frequency (VAF) $\geq$ 8% and normal VAF $\leq$ 5%, in addition to the significance tests already performed by the programs. (2) A list of all candidate mutations was collated requiring that a mutation was either seen in the high-confidence somatic set, the set of variants from the previously published study[14] or from the lifted-over set of variants from the TCGA AML paper[6]. (3) Mutations from the overall set were kept if: (a) the overall number of calls in the paired samples was not more than twice the number of high-confidence somatic calls; (b) the tumour-only frequency for the calls was less than 50% greater than the number of calls in the paired samples; (c) the mutation was seen in list of the previous study[14] or TCGA dataset[6]. (4) High-confidence somatic mutations were kept regardless.

The data from the two genotypers were combined along with FLT3-ITD calls from Pindel[54]. Comparing our variant lists from whole-exome sequencing amd custom-capture validation sequencing, we noticed—similar to others[55]—that low allele frequency C-to-A variants (<15%) tended to have poor concordance (7.7%; data not shown) between the initial run and the technical validation run. These variants were removed in these data, along with a curated 'blacklist' (Supplementary Table 15) of known problematic variants and/or genes, including mitochondrial DNA variants. In addition, all variants that were seen in a cumulative list of normal samples from Beat AML at a frequency greater than 1% were removed. Cumulatively, of this set, 94% of covered SNVs were validated with 82% of indel calls also being confirmed with validation sequencing. Manual review was then carried out in the following steps. (1) The addition back of all flagged rows from the previous study[14]. (2) The review of all TCGA-flagged rows for VAF pattern that matched or did not match with known drivers in the same specimen. Some TCGA

variants were added back based on convincing VAF pattern and known pathogenic role, other TCGA variants were kept excluded based on a VAF pattern that was unlike known drivers in the same specimens. (3) Other variants were added back based on other specimens that had the same variant that were still on the include list, and if the VAF pattern looked convincing for inclusion. (4) All genes from the previous study[14] with only frameshift and/or nonsense variants were manually reviewed and missense mutations were manually removed. (5) Genes and/or variants that were on both the include and exclude lists were manually reviewed and they were removed if they were C-to-A with over 15% VAF, they did not validate and/or the VAF pattern was unlike known drivers in same specimen. (6) Further review of all genes in the summary sheet with cohort frequency of 8 or more (1% of more). Any that were not familiar from knowledge of AML literature were manually reviewed for VAF patterns that did or did not match known drivers within the same specimens. Those that did not match were manually removed.

After this manual review, additional curated mutations from the UnifiedGenotyper run were added back in along with a curated set of variants from tumour-only patients for genes from the previous study[14]. The tumour-only AML samples used in another study (H.Z. et al., manuscript submitted) were removed and used for that study.

Detection of internal FLT3-ITD and *NPM1* mutations. A subset of samples was tested for FLT3-ITD and *NPM1* mutation status using an internally run PCR assay and capillary electrophoresis. Genomic DNA (gDNA) was extracted from fresh blood or bone marrow aspirates of patients with AML and was used to detect the presence or absence of FLT3-ITD and *NPM1* 4-bp insertion mutations[56,57]. Primers for *FLT3* spanned approximately 330 bp to include the common internal duplication site[56]. Primers for *NPM1* spanned approximately 170 bp to cover the clustered multiple insertional or indel sites[57,58]. Primers were HPLC-purified by the manufacturer. The multiplex PCR reaction solution[59] consisted of 100 ng gDNA, 10 pmol of the respective forward and reverse primers for *FLT3* and *NPM1*, 25 mmol l^{-1} $MgCl_2$, 2.5 mmol l^{-1} dNTPs, 5 μl 10× PCR buffer, 0.2 μl AccuTaq DNA polymerase and water in a total volume of 50 μl. The PCR conditions were: initial denaturing for 5 min at 94 °C, followed by 30 cycles at 94 °C for 30 s, 56 °C for 45 s and 72 °C for 30 s with a final cycle of 10 min at 72 °C. The PCR products were diluted 1:10 and analysed by capillary electrophoresis on a QIAxcel high-resolution DNA cartridge according to the manufacturer's protocol. Forward primer *FLT3*: 5′-AGCAATTTAGGTATGAAAGCCAGCTA-3′; reverse primer *FLT3*: 5′-CTTTCAGCATTTTGACGGCAACC-3′. Forward primer *NPM1*: 5′-GTTTCTTTTTTTTTTTTTTCCAGGCTATTCAAG-3′; reverse primer *NPM1*: 5′-CACGGTAGGGAAAGTTCTCACTCTGC-3′.

Derivation of FLT3-ITD and *NPM1* consensus calls. Consensus FLT3-ITD and *NPM1* mutation calls found in the clinical summary table (Supplementary Table 5) were determined by comparing the internal capillary PCR test (internal; according to the methods described in 'Detection of internal FLT3-ITD and *NPM1* mutations'), the Clinical Laboratory Improvement Amendments/College of American Pathology (CLIA/CAP) laboratory run test (Sequenome, GeneTrails, Foundation Medicine, Genoptix, Illumina). The internal test was used for the sample consensus call when available, as it was performed on the exact sample that was used for further ex vivo drug sensitivity assays. Where discordance existed between the internal test and the CLIA laboratory test results, the sample was flagged for manual review. The trace file for the internal test was visually inspected and if discordance with the CLIA/CAP test results persisted, the whole-exome sequencing data were then used to help to determine the consensus call.

Derivation of CCAAT enhancer binding protein α (*CEBPA*) biallelic consensus calls. N-terminal and C-terminal mutations have been described to occur on opposing alleles and patients with *CEBPA* biallelic mutations have been shown to fall into a favourable risk category[60]. Patients were scored positive for biallelic *CEBPA* mutation if described in the clinical notes as biallelic or double-positive. Patients were also scored as *CEBPA* biallelic if both N-terminal and C-terminal mutations were identified in the whole-exome sequencing data.

RNA sequencing and data processing. All samples were sequenced using the Agilent SureSelect Strand-Specific RNA Library Preparation Kit on the Bravo robot (Agilent). In brief, poly(A)$^+$ RNA was chemically fragmented. Double-stranded cDNAs were synthesized using random hexamer priming with 3′ ends of the cDNA adenylated, after which indexed adaptors were ligated. Library amplification was performed using three-primer PCR using a uracil DNA glycosylase addition for strandedness. Libraries were validated with the Bioanalyzer (Agilent) and combined to run 4 samples per lane, with a targeted yield of 200 million clusters. Combined libraries were denatured, clustered with the cBot (Illumina) and sequenced on the HiSeq 2500 using a 100-cycle paired-end protocol. In addition to the AML samples, there was also a sample of purified CD34 molecule (CD34)$^+$ cells from healthy control bone marrow, which was included in each sample group (for a total of 12 times sequencing this control RNA). This control served as both a healthy control and a quality check on inter-group batch effects. In addition, 21 individual healthy bone marrow samples were also included, two of which were CD34-selected (17-00053 and 17-00056) with the other 19 being whole mononuclear bone marrow cells from healthy donors.

Workflows for processing and analysis of RNA-sequencing data to generate gene counts and gene fusions for each sample are shown on our online browser (http://vizome.org/additional_figures_BeatAML.html) with processed gene expression values for each specimen listed in Supplementary Tables 8, 9. For each flowcell and each sample, the FASTQ files were aggregated into single files for read 1 and read 2 (if not already done by the sequencing core). During this process, these reads were trimmed by 3 on the 5′ end and 5 on the 3′ end. Alignment of reads was performed using the subjunc aligner (version 1.5.0-p2)[61]. BAM files obtained from subjunc were used as inputs into featureCounts (version 1.5.0-p2)[62] and gene-level read counts were produced. For a reference genome, the GRCh37 build provided by the Broad as part of the GATK bundle was used. Gene assignments were based on the Ensembl build 75 gene models on GRCh37. The following parameters were used for the software:

subjunc -i /path/to/reference/ -u -r fastq1 -R fastq2 -o outputBAMFilename -I 5 -T 7 -d 50 -D 600 -S fr and featureCounts -a Homo_sapiens.GRCh37.75.gtf -o output -F GTF -t exon -g gene_id -s 2 -C -T 10 -p -P -d 50 -D 600 -B BAM_files.

The data were collated from featureCounts matrices and all genes with no counts across the samples were excluded. Genes with duplicate gene symbols and those for wich the counts were <10 for 90% or more of the samples were additionally removed before normalization similar to the approach suggested for weighted gene correlation network analysis (WGCNA)[63]. Samples for which the median expression was less than 2 standard deviations below the average were removed from the dataset ($n = 10$). Normalization was performed using the conditional quantile normalization procedure[64], which produced GC-content-corrected $\log_2$ reads per kilobase per million mapped reads (RPKM) values. This procedure produces both offsets to be used in conjunction with edgeR as well as a matrix of $\log_2$-normalized RPKM values for clustering.

In addition, the subjunc BAM files were processed using the RNA-sequencing genotyping protocol (as of GATK version 3.3), which was similar to the whole-exome sequencing protocol described in the 'Whole-exome sequencing data processing' section, including the following steps for each sample: (1) MarkDuplicates; (2) SplitNCigarReads; (3) RealignerTargetCreator/IndelRealigner; (4) base quality score recalibration. The resulting BAM files were used to produce RNA genotypes using the UnifiedGenotyper for the purposes of quality control and ethnicity estimation.

Gene fusion data were additionally generated using the TopHat-Fusion (version 2.0.14) program using default parameters[65].

Coexpression network formation. We formed coexpression modules using the WGCNA procedure on the RNA-sequencing data from the 'RNA sequencing and data processing' section. All RNA-sequencing samples were used to form the set of modules. Owing to the heterogeneity of the clinical expression data, we generated 'signed hybrid' networks using the 'bicor' correlation[66], setting the maximum proportion of outliers to 0.1. We ran the procedure multiple times, varying several parameters to choose the most relevant set for further analysis. The WGCNA procedure was run on datasets formed from the top 2,000 and 5,000 most variable genes. For each dataset, we set the 'power' variable to either 2 or 3. For each of these runs, we varied the module detection parameters of dynamicTreeCut[63], namely the deepSplit parameter was set to 0 or 2 and the pamStage parameter was set to TRUE or FALSE. For each of these sets of modules we computed a series of module quality statistics[67], mean correlation, mean adjacency, mean maximum adjacency ratio (MAR), mean correlation with the module eigengene (KME), proportion of variability explained and the mean cluster coefficient. Significance of modules was determined by computing a z score of each of these values relative to the mean and standard deviation of those from 100 random assignment of modules. We chose the set of modules to use in our analyses as those that were most correlated with the 'specimentSpecificDx' using module quality as a tie-breaker. The analysis set of modules was chosen to be the version using the 5,000 most variable genes, power set to 2 and modules formed using deepSplit = 2 and pamStage = F. Of this set of modules, only the grey module did not have a summary z statistic (median across the four density measures) of at least 2. In addition, after correcting the data using the estimated principal components[68], the module structure did not change appreciably (data not shown).

Quality control. The UnifiedGenotyper runs for both the whole-exome sequencing and RNA-sequencing data were combined into a single VCF file using the GATK CombineVCFs functionality. This combined VCF file was converted to a GDS file using SNPRelate (version 1.12.2)[69]. Note, the version is the most recent version as several versions were used across the entire project. The overall similarity of the genotypes of each pair of samples was computed, termed identity by state (IBS) and hierarchical clustering was performed using one minus this similarity. From this clustering and visualization we had devised hard cut-offs for further inspection based on the types of data being compared. For instance, samples not meeting the specified IBS thresholds (DNA–DNA = 0.9; RNA–RNA = 0.83;

DNA–RNA = 0.89) were subject to manual review. On the basis of the dendrogram structure as well as the clinical/laboratory information, samples were either excluded, assigned to a different patient ID or in rare cases assigned to a different sample. It was observed that bone marrow transplants between sample collections produced a noticeable but milder effect in these dendrograms and such samples were flagged for removal from the RNA-sequencing analysis and for treatment as tumour-only samples in the whole-exome sequencing analysis as described in 'Whole-exome sequencing variant detection'.

Fusion annotation for analysis. Fusions calls were determined from a consensus of three datatypes, a specific diagnosis categorization at the time of sample acquisition, current set of clinical karyotypes and fusions detected in RNA-sequencing data by TopHat-Fusion. All sources were limited to the same set of known fusions: *RUNX1-RUNX1T1*, *CBFB-MYH11*, *MLLT3-KMT2A*, *DEK-NUP214*, *GATA2-MECOM* and *PML-RARA*. It was determined that the RNA-sequencing calls did not provide additional resolution in detecting these known fusions and was not performed on all the samples, so the consensus was limited to the clinical karyotype calls as well as the specific diagnosis categorization (which was determined based on karyotype and other cytogenetic clinical tests). Overall, the calls were based on the karyotype data except in 10 cases; for 3 cases the karyotype and diagnosis was sufficiently complex to warrant a separate 'complex' categorization. The remaining 7 of these cases were set to the specific diagnosis classification. It should be noted that there was additional support from the RNA-sequencing data for several of these cases.

Ethnicity. The combined RNA and whole-exome sequencing VCF files from the 'Quality control' step were merged with a set of Hapmap genotypes[70] lifted over to build 37. The SNPRelate package was used to convert the VCF to GDS, perform linkage disequilibrium (LD) pruning using an LD threshold of 0.2, MAF cut-off of 0.05 and allowing a missing rate of 0.3 and calculation of the principal components. Previously published methodology[71] was used to assign admixture proportions relative to the HapMap samples using the principal components. Each sample was assigned to an ethnicity group based on the group with the maximum admixture proportion. If the maximum was 50% or less, we labelled it 'Admixed'. As we had observed previously that the clustering of ethnicities for RNA-sequencing samples are more diffuse than exome sequencing, we assigned the final inferred ethnicity to each patient based on the distinct whole-exome sequencing calls if available, deferring to RNA sequencing only if not available. If multiple exome sequencing samples were present with discrepant calls, the data of the patient were manually reviewed. The only patient for which this occurred was patient 4043, a self-identified Hispanic who had two RNA samples and an exome sequencing sample inferred as 'White' and one exome sample inferred to be 'HispNative'. Only the 'White' call for the exome sequencing had an admixture proportion over 0.5. The patient was kept consistent with the self-identification and labelled as 'HispNative'.

Sex. For DNA, coverage was first computed over the Y chromosome and the counts for each sample were added up and $\log_{10}$-transformed (after adding 1 to all the counts). *k*-means clustering was used to assign samples to two clusters with the cluster with the lower mean labelled as the 'Female' cluster.

For RNA sequencing, counts were converted to counts per million after applying the Trimmed Mean of M scaling normalization[72]. A set of 28 genes were chosen to successfully discriminate the genders using DE analysis over multiple studies (data not shown) and were used in conjunction with *k*-means clustering to form two clusters. The 'Female' cluster was labelled based on high *XIST* expression.

ELN 2017 classification. This procedure is based on the categorization in table 5 of the 2017 ELN update paper[5]. Karyotypes in the clinical file were first cleaned and parsed into clones or subclones and distinct abnormalities using standard conventions[73]. The current representation was corrected for nomenclature type (for example, 'idem' versus 'sl') in a basic manner. For instance, ambiguous events, such as chromosomal loss (for example, −15), were not corrected for whether the preceding clone had a counteracting gain. Also, additional '+' or '−' symbols in conjunction with valid karyotype operators in a separate clone (for example, +del(12)(q?15) or −del(12)(q?15))) were treated separately with gains (+) being kept in the unique count of events and losses (−) being removed.

Abnormalities were first checked for the following categories: (1) *RUNX1-RUNX1T1*; (2) *CBFB-MYH11*; (3) *MLLT3-KMT2A*; (4) *DEK-NUP214*; (5) *KMT2A-**; (6) *BCR-ABL1*; (7) *GATA2-MECOM*; (8) −5/del(5q); (9) −7; (10) −17. Categories 1–8 were further considered to be WHO recurrent fusions.

The number of unique abnormalities (across clones) was then computed. Whether or not a karyotype was considered to be polyploid was also recorded (at least 60 chromosomes or '($\geq$ 3)n' or label). *NPM1*, FLT3-ITD and biallelic *CEBPA* were derived from consensus calls. FLT3-ITD allelic ratios were determined only for the samples with an internal assay. The MAF values of the internal assay were converted to a ratio using the formula MAF/(1 − MAF). *RUNX1*, *ASXL1* and *TP53* were derived from the clinical genotypes. Abnormal 17 calls were manually curated from the karyotype data and clinical genotype calls.

The determination of ELN 2017 categories proceeds by assigning true/false/not available values to one or more of the five columns (three ELN and two ambiguous) in the following manner. (1) 'isFavourable' is considered true if a sample has at least one of the following: (a) *RUNX1-RUNX1T1*; (b) *CBFB-MYH11*; (c) positive *NPM1* and negative FLT3-ITD; (d) positive *NPM1* and positive FLT3-ITD with allelic ratio <0.5; (e) biallelic *CEBPA*. (2) 'isFavourableOrIntermediate' when *NPM1* is positive and FLT3-ITD is positive but the allelic ratio is not available. (3) 'isAdverse' is considered true if a sample has at least one of the following: (a) *DEK-NUP214*; (b) *KMT2A-**; (c) *BCR-ABL1*; (d) *GATA2-MECOM*; (e) −5/del(5q); (f) −7; (g) −17; (h) abn_17; (i) three or more abnormalities and no WHO recurrent fusions; (j) one monosomy (autosomal) and at least one additional abnormality except for *CBFB-MYH11*; (k) positive *RUNX1* or *ASXL1* and not considered to be 'isFavourable' or 'isFavourableOrIntermediate'; (l) positive *TP53*; (4) 'isIntermediate' is considered true if a sample has at least one of the following: (a) *MLLT3-KMT2A*; (b) *NPM1* is positive and FLT3-ITD is positive with allelic ratio ≥0.5; (c) *NPM1* is negative and FLT3-ITD is negative or has a low allelic ratio (<0.5); (d) at least one abnormality and is not considered 'isFavourable' or 'isAdverse'; (5) 'isIntermediateOrAdverse' when *NPM1* is negative and FLT3-ITD is positive without an allelic ratio. Calls were annotated as 'not available' in the absence of FLT3-ITD or *NPM1* calls.

Samples for which the specific diagnosis at inclusion indicated 'acute promyelocytic leukaemia with t(15;17)(q22;q12)' were automatically set to 'Favourable'. Any overlaps in the categories were resolved based on manual expert review.

Ex vivo functional drug screens. Ex vivo functional drug screens were performed on freshly isolated mononuclear cells from AML samples. In brief, 10,000 cells per well were arrayed into three, 384-well plates containing 122 small-molecule inhibitors. This panel contained graded concentrations of drugs with activity against two-thirds of the tyrosine kinome as well as other non-tyrosine kinase pathways, including mitogen-activated protein kinases (MAPKs), the pathway involving phosphatidylinositol-4,5-bisphosphate 3-kinase, AKT serine/threonine kinase 1 and mechanistic target of rapamycin kinase (PIK3C–AKT–MTOR); protein kinase AMP-activated (AMPK, also known as PRKAA1), ATM serine/threonine kinase (ATM), Aurora kinases, calcium/calmodulin-dependent protein kinases (CAMKs), cyclin-dependent kinases (CDKs), serine/threonine protein kinase 3 (GSK3), IκB kinase (IκK), cAMP-dependent protein kinase (PKA), protein kinase C (PKC), polo-like kinase 1 (PLK1) and RAF proto-oncogene serine/threonine kinase (RAF). In addition, the library contained small-molecule inhibitors with activity against the BCL2 family, bromodomain containing 4 (BRD4), Hedgehog, heat shock protein 90 (HSP90), NOTCH/γ-secretase, proteasome, survivin, signal transducer and activator of transcription 3 (STAT3), histone deacetylase (HDAC), and WNT/β-catenin. Drug plates were created using inhibitors purchased from LC Laboratories and Selleck Chemicals and master stocks were reconstituted in dimethyl sulfoxide (DMSO) and stored at −80 °C. Master plates were created by distributing a single agent per well in a seven-point concentration series, created from threefold dilutions of the most concentrated stock resulting in a range of 10 μM to 0.0137 μM for each drug (except dasatinib, ponatinib, sunitinib and YM-155, which were plated at a concentration range of 1 μM to 0.00137 μM). DMSO-control wells and positive-control wells containing a drug combination of flavopiridol, staurosporine and velcade were placed on each plate, with the final concentration of DMSO ≤0.1% in all wells. Daughter plates were created using a V&P Scientific 384-well pin tool head operated by the Caliper Sciclone ALH 3000 and equipped with 0.457-mm diameter, 30-nl, slotted stainless-steel pins (FP1NS30). Daughter and destination plates were sealed with pealable thermal seals using a PlateLoc thermal sealer. Destination plates were stored at −20 °C for no more than three months and thawed immediately before use. Primary mononuclear cells were plated across single-agent inhibitor panels within 24 h of collection. Cells were seeded into 384-well assay plates at 10,000 cells per well in Roswell Park Memorial Institute (RPMI) 1640 medium supplemented with fetal bovine serum (FBS) (10%), L-glutamine, penicillin–streptomycin, and β-mercaptoethanol (10–4 M). After three days of culture at 37 °C in 5% CO_2, MTS reagent (CellTiter96 AQueous One; Promega) was added, the optical density was measured at 490 nm, and raw absorbance values were adjusted to a reference blank value and then used to determine cell viability (normalized to untreated control wells).

Ex vivo functional drug screen data processing. A workflow in which the data normalization, curve fit parameters and quality assurance/quality control steps are summarized can be found on our online browser (http://vizome.org/additional_figures_BeatAML.html) with processed drug response data for each specimen listed in Supplementary Table 10. A given sample was run on one or more panels and within each panel, the majority of drugs were run without within-panel replicates. Two steps were performed to harmonize these data before model fitting.

First, a 'curve-free' AUC (integration based on fine linear interpolation between the seven data points themselves) was calculated for those runs with within-panel replicates after applying a ceiling of 100 and a floor of 0 for the normalized viability. The maximum change in AUC among the replicates was noted and those runs with differences >100 were removed.

Second, the remaining within-plate replicates had their normalized viability averaged subject to a ceiling of 100 and floor of 0. An additional set of 'curve-free' AUCs was computed for sample–inhibitor pairs run on multiple panels. The maximum change in AUC among the across-panel replicates was noted and those runs with differences >75 were removed.

At this point, the within- and across-plate replicates for the normalized viability were averaged together and a ceiling of 100 was applied. From the steps above, the floor was already at 0. On the basis of the methodology used in our prior drug-combination study[74], a probit regression was fit to all possible run groups using the model: (normalized viability / 100) ~ $1 + \log_{10}$(concentration). For all groups there were $n = 7$ dose–response measurements.

The summary analyses of curve fit were inspected and cut-offs were devised removing all runs with an Akaike information criteria (AIC) > 12 and deviance >2. For inhibitors that were run using multiple concentration ranges, only the most-recent concentration range was kept. Finally, these data were compared to the AUC values from third-order polynomial fits. Those runs that were discrepant in terms of sensitive or resistant calls were manually reviewed as subject to removal.

Ex vivo functional drug screen analysis. For all drug analyses that required a call of sensitivity or resistance (for example, the gene expression signatures), sensitivity or resistance was determined by the lowest and highest 20% of the AUC values for each drug.

Correlations between drugs in families. For each inhibitor in the study, available data on targets of the inhibitors were pulled from a variety of online resources and published studies, many of which were aggregated in the Cancer-Targetome[75,76]. Activity of each inhibitor for targets was then distilled into a five-tier system to afford comparison across drugs with differing degrees of potency and/or for which differing assays were used to measure drug and/or target activity. Well-represented genes, gene families and pathways were then filtered for drugs that have activity in the top 3 tiers for one or more member of the gene family or pathway. These lists were then manually curated to generate the final list of high-confidence drug target families (shown in Supplementary Table 11). For each inhibitor assigned to at least one drug target family, the Pearson's correlation was computed against all other drugs assigned to at least one drug target family for the AUC values of all available samples shared between the two drugs.

Correlations between drugs and samples. Drugs were first filtered to require greater than two hundred samples per drug. Additional samples were removed accordingly to allow correlations to be computed between all present samples using available AUC data and between all drugs.

Summary drug response scores. For each patient sample, a binary score (1/0) was used for each drug based on the same threshold as for the gene signatures (for example, sensitivity or resistance was determined by the lowest and highest 20% of the AUC values for each drug). Individual scores were computed for resistance and sensitivity separately and presented as the proportion over all screened drugs for each patient sample.

Expression analysis and integration with ex vivo functional drug screen. For all the below analyses the earliest sample was chosen for each patient.

Expression heat map. The top 2,000 most variable genes were extracted. The expression values were centred and scaled across patients and complete-linkage hierarchical clustering was performed using the ComplexHeat map R package[77].

Sensitive or resistant differential expression. For each drug, it was required that at least three sensitive and three resistant samples using the 20%/20% criteria outlined in the 'Ex vivo functional drug screen analysis' section. Patient samples were limited to those labelled as sensitive or resistant. Next, genes were limited based on their expression, for which at least half the patients used for analysis had to have greater than one count per million (an approach suggested in the limma users manual)[78]. The normalized expression as in the data described in the 'RNA sequencing and data processing' section with the chosen samples and genes was used for differential expression analysis. Because the data had not been batch-corrected at this point, surrogate variable analysis (SVA)[79] was used to infer covariates for correcting out technical confounders. Next, the linear model fitting for each gene was performed using the limma-trend approach[80], testing whether the average expression was different between resistant and sensitive correcting for the SVA covariates. Genes with Benjamini–Hochberg[38] FDR values of less than 0.05 were kept for the cluster analysis. The expression matrix was corrected with respect to the estimated surrogate variables for consistency with the differential expression procedure using fSVA[81] and MCLUST[82] was used to determine optimal number of clusters and parameterization. The results were then visualized using a CLUSPLOT[83], which displays the clustering results with respect to the first two principal components of the gene expression for the kept genes.

Mutation analysis and integration with ex vivo functional drug screen. For all the below analyses in which groups of samples were compared, the earliest sample was chosen for each patient.

TCGA comparison. The lifted-over TCGA variants from the 'Whole-exome sequencing variant detection' section were annotated using the VEP from Ensembl build 83, filtered for protein-altering and splice site variants and our 'blacklist' was applied to ensure the variant sets were comparable.

Co-occurrence and mutual exclusivity. Only mutations seen in at least 10 patients were kept. The DISCOVER[41] method was used to determine significant mutual exclusivity and co-occurrence. A plot of the co-occurrences was generated using corrplot[84] with the odds ratio of the pairwise co-occurrence used to colour and scale the circle sizes.

Association between mutations and drugs. For each mutated gene in the exome sequencing samples and each recurrent fusion (counting FLT3-ITD as a distinct entity from other *FLT3* mutants), we determined all available (at least 5 patients) pairwise and three-way co-occurrence sets. For each drug and each valid set of genes (from one to three genes), we fitted a linear model with AUC as the response and examined the linear contrast (that is, two-sided Student's t-test) comparing the AUC of the gene(s) to the appropriate negative. For example the average AUC of the *FLT3*, *DNMT3A* and *NPM1* mutants would be compared to average AUC of the samples negative for all three genes. FDR was computed using the Benjamini–Hochberg method over all the drugs.

For the ibrutinib and entospletinib comparisons, the presence and absence of the three genes or mutations: *NPM1*, FLT3-ITD and *DNMT3A* was collapsed into levels of a single factor. The corresponding single-factor ANOVA was carried out with the 'triple-negative' category set as the reference. Significance of the P values of each coefficient was compared to the Bonferroni-corrected 0.05 level.

For the JAK-family analysis, the AUC values were pooled for the four JAK inhibitors (CYT387, tofacitinib (CP-690550), JAK inhibitor I, ruxolitinib (INCB018424)) for each gene mutation set (*BCOR*, *BCOR* and DNMT3A, *BCOR* and *RUNX1*, *BCOR* and *SRSF2*). The contrast of the difference between *BCOR* and *RUNX1* samples and the average of the other three mutation groups was tested.

Integration of both mutation and RNA sequencing with ex vivo functional drug screen. Mutations (0/1 score) and the module Eigengenes from the WGCNA analysis were used separately and combined together in regression models with coefficients selected using the LASSO approach[40] as implemented in glmnet[85]. For each data type and the combination, only drugs with at least 200 patients samples were tested. The three datasets were initially randomly separated into training (75%) and test (25%) sets. Similar to a previous approach[86], a bootstrap aggregation approach was used in which the 1,000 bootstraps of the training dataset were generated and for each one, the LASSO was trained using tenfold cross-validation. Predictions were formed for the test dataset over these bootstrap models and the predicted AUC was averaged. R^2 values were computed for these aggregated predictions relative to the test AUC values. As performance was seen to be dependent on the initial split between test and training, we repeated the entire process 100 times, recording the mean and standard deviation of the R^2 value as well as the count non-zero coefficients

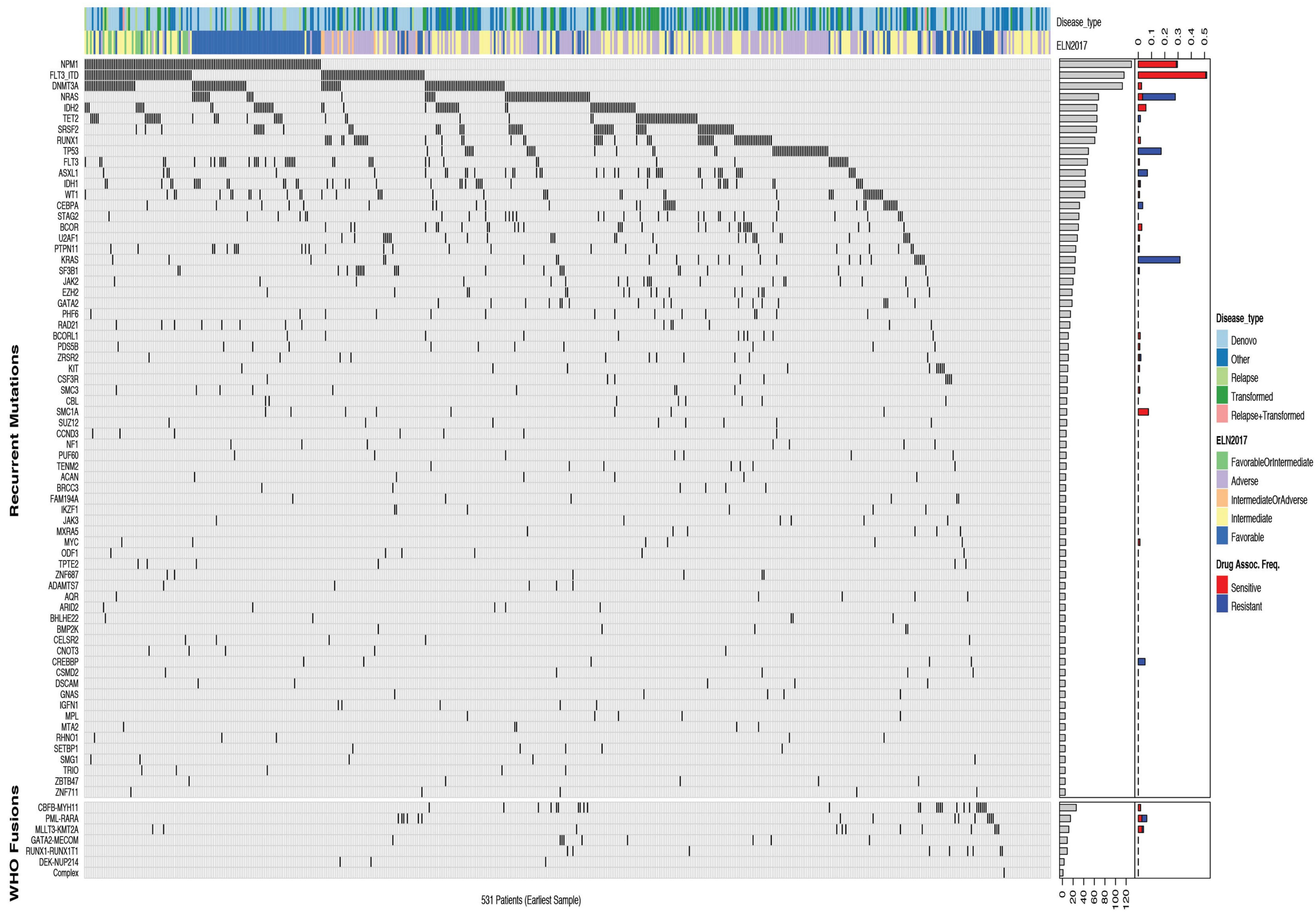
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

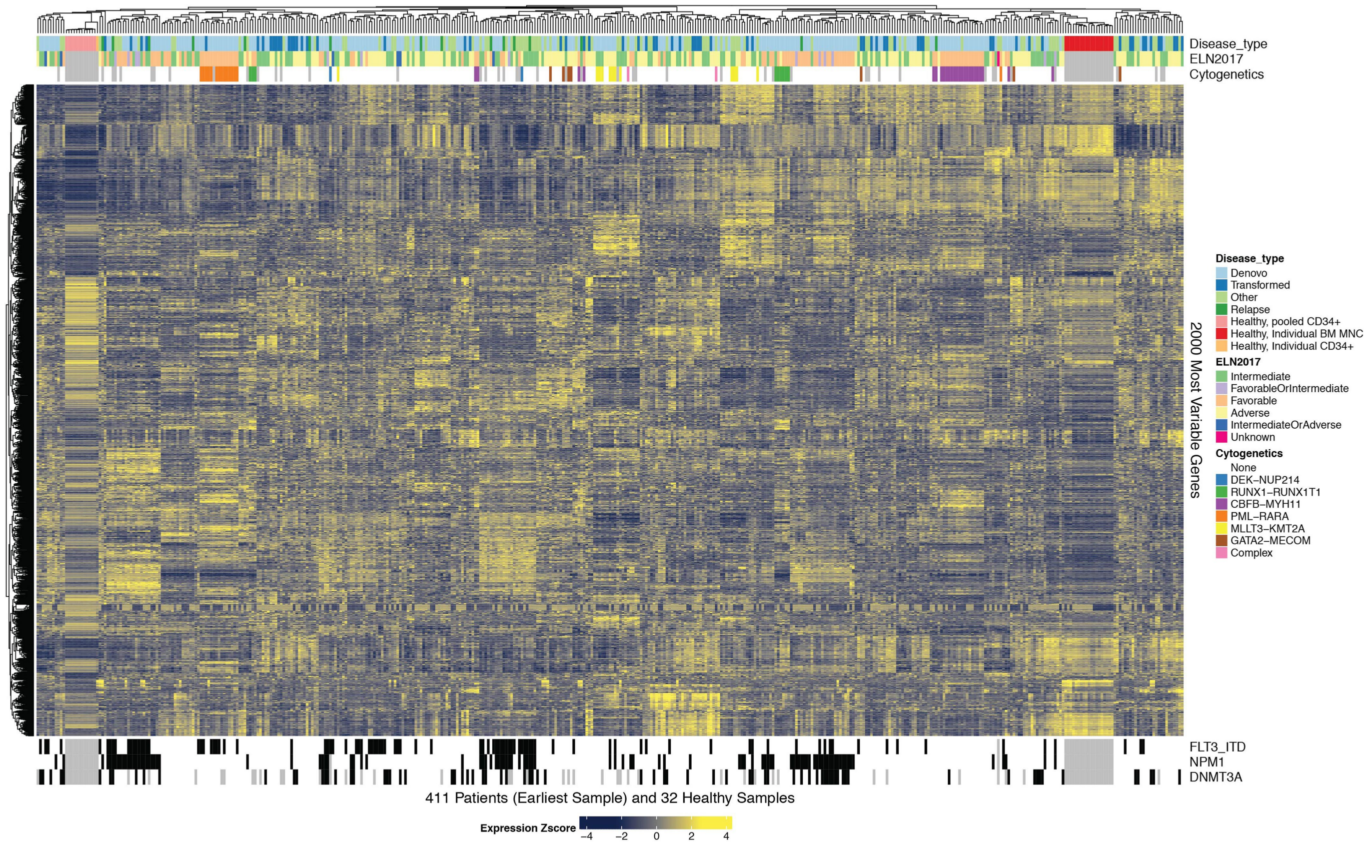
All raw and processed sequencing data, along with relevant clinical annotations, have been submitted to dbGaP and Genomic Data Commons. The dbGaP study ID is 30641 and accession ID is phs001657.v1.p1. The raw data for clinical annotations, variant calls, gene expression counts and drug sensitivity that underlie Figs. 1–3 and Extended Data Figs. 1–9 are provided as Source Data. In addition, all data can be accessed and queried through our online, interactive user interface, Vizome, at http://www.vizome.org/.

42. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
43. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
44. Huntley, M. A. et al. ReportingTools: an automated result processing and presentation toolkit for high-throughput genomic analyses. *Bioinformatics* **29**, 3220–3221 (2013).
45. Buffalo, V. qrqc: Quick Read Quality Control. R package version 1.22.0 http://github.com/vsbuffalo/qrqc (2012).
46. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
47. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
48. McLaren, W. et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
49. Memorial Sloan Kettering. vcf2maf. version 1.6.6 https://github.com/mskcc/vcf2maf/ (2016).
50. Koboldt, D. Release note for Varscan version 2.4.1. https://github.com/dkoboldt/varscan/blob/master/VarScan.v2.4.1.description.txt (2015).
51. Memorial Sloan Kettering. maf2vcf. version 1.6.6 https://github.com/mskcc/vcf2maf/ (2016).
52. Zhao, H. et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).

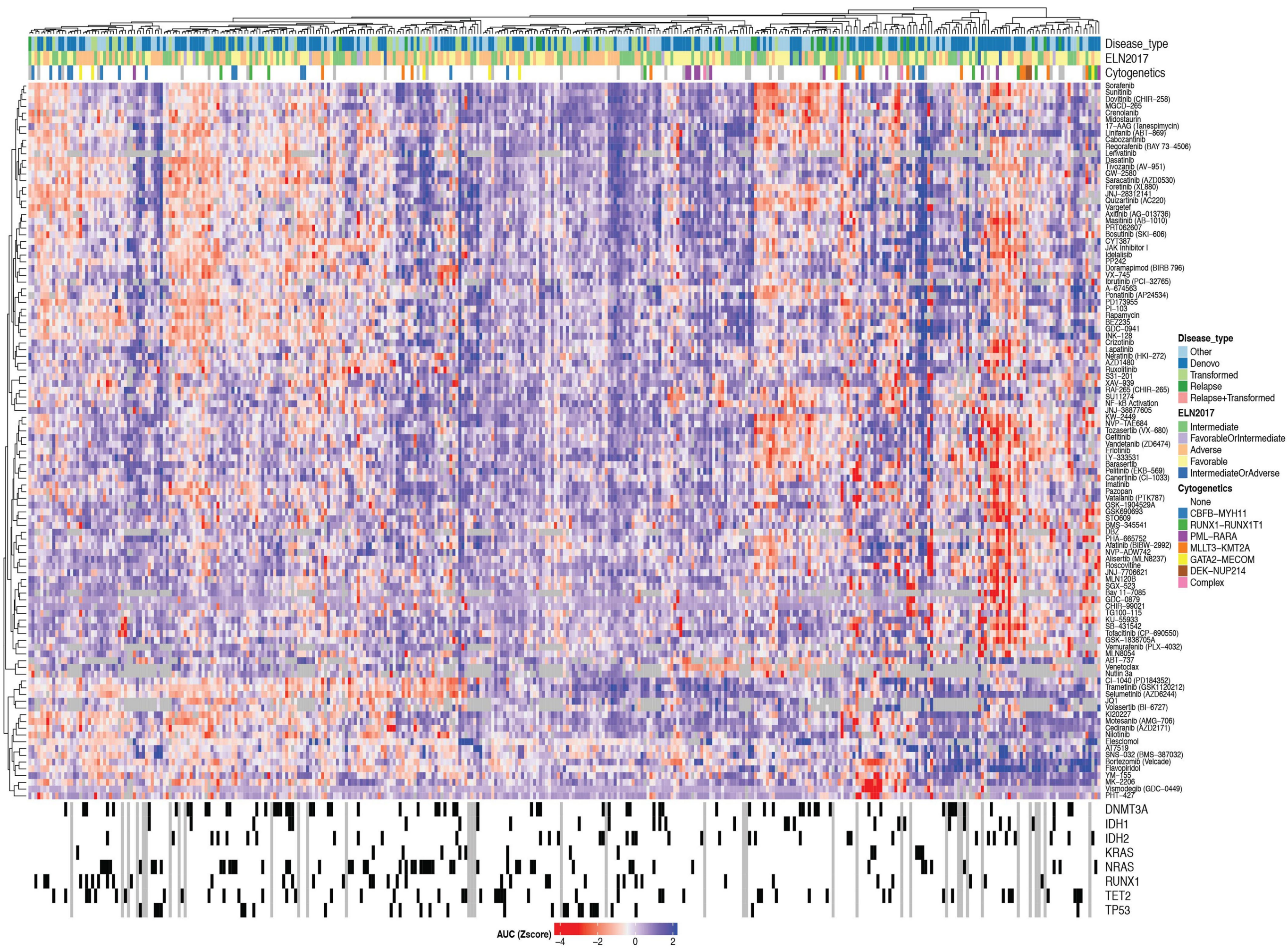
53. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
54. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
55. Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
56. Kottaridis, P. D. et al. The presence of a *FLT3* internal tandem duplication in patients with acute myeloid leukemia (AML) adds important prognostic information to cytogenetic risk group and response to the first cycle of chemotherapy: analysis of 854 patients from the United Kingdom Medical Research Council AML 10 and 12 trials. *Blood* **98**, 1752–1759 (2001).
57. Döhner, K. et al. Mutant nucleophosmin (*NPM1*) predicts favorable prognosis in younger adults with acute myeloid leukemia and normal cytogenetics: interaction with other gene mutations. *Blood* **106**, 3740–3746 (2005).
58. Falini, B., Nicoletti, I., Martelli, M. F. & Mecucci, C. Acute myeloid leukemia carrying cytoplasmic/mutated nucleophosmin (NPMc^{+} AML): biologic and clinical features. *Blood* **109**, 874–885 (2007).
59. Huang, Q. et al. A rapid, one step assay for simultaneous detection of *FLT3*/ITD and *NPM1* mutations in AML with normal cytogenetics. *Br. J. Haematol.* **142**, 489–492 (2008).
60. Wouters, B. J. et al. Double *CEBPA* mutations, but not single *CEBPA* mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113**, 3088–3091 (2009).
61. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108 (2013).
62. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
63. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
64. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).
65. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**, R72 (2011).
66. Langfelder, P. & Horvath, S. Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* **46**, 1–17 (2012).
67. Langfelder, P., Luo, R., Oldham, M. C. & Horvath, S. Is my network module preserved and reproducible? *PLOS Comput. Biol.* **7**, e1001057 (2011).
68. Parsana, P. et al. Addressing confounding artifacts in reconstruction of gene co-expression networks. Preprint at https://www.biorxiv.org/content/early/2017/10/13/202903 (2017).
69. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).
70. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
71. Zheng, X. & Weir, B. S. Eigenanalysis of SNP data with an identity by descent interpretation. *Theor. Popul. Biol.* **107**, 65–76 (2016).
72. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
73. Slovak, M. L., Theisen, A. & Shaffer, L. G. in *The Principles of Clinical Cytogenetics* (eds Gersen, S. L. & Keagle, M. B.) 23–49 (Springer, New York, 2013).
74. Kurtz, S. E. et al. Molecularly targeted drug combinations demonstrate selective effectiveness for myeloid- and lymphoid-derived hematologic malignancies. *Proc. Natl Acad. Sci. USA* **114**, E7554–E7563 (2017).
75. Davis, M. I. et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).
76. Blucher, A. S., Choonoo, G., Kulesz-Martin, M., Wu, G. & McWeeney, S. K. Evidence-based precision oncology with the cancer targetome. *Trends Pharmacol. Sci.* **38**, 1085–1099 (2017).
77. Gu, Z., Eils, R. & Schlesner, M. Complex heat maps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
78. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
79. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, e161 (2007).
80. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
81. Parker, H. S., Corrada Bravo, H. & Leek, J. T. Removing batch effects for prediction problems with frozen surrogate variable analysis. *PeerJ* **2**, e561 (2014).
82. Fraley, C. & Raftery, A. E. Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. *J. Classif.* **20**, 263–286 (2003).
83. Pison, G., Struyf, A. & Rousseeuw, P. J. Displaying a clustering with CLUSPLOT. *Comput. Stat. Data Anal.* **30**, 381–392 (1999).
84. Wei, T. et al. corrplot: Visualization of a Correlation Matrix. R package version 0.84 https://github.com/taiyun/corrplot (2017).
85. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
86. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).



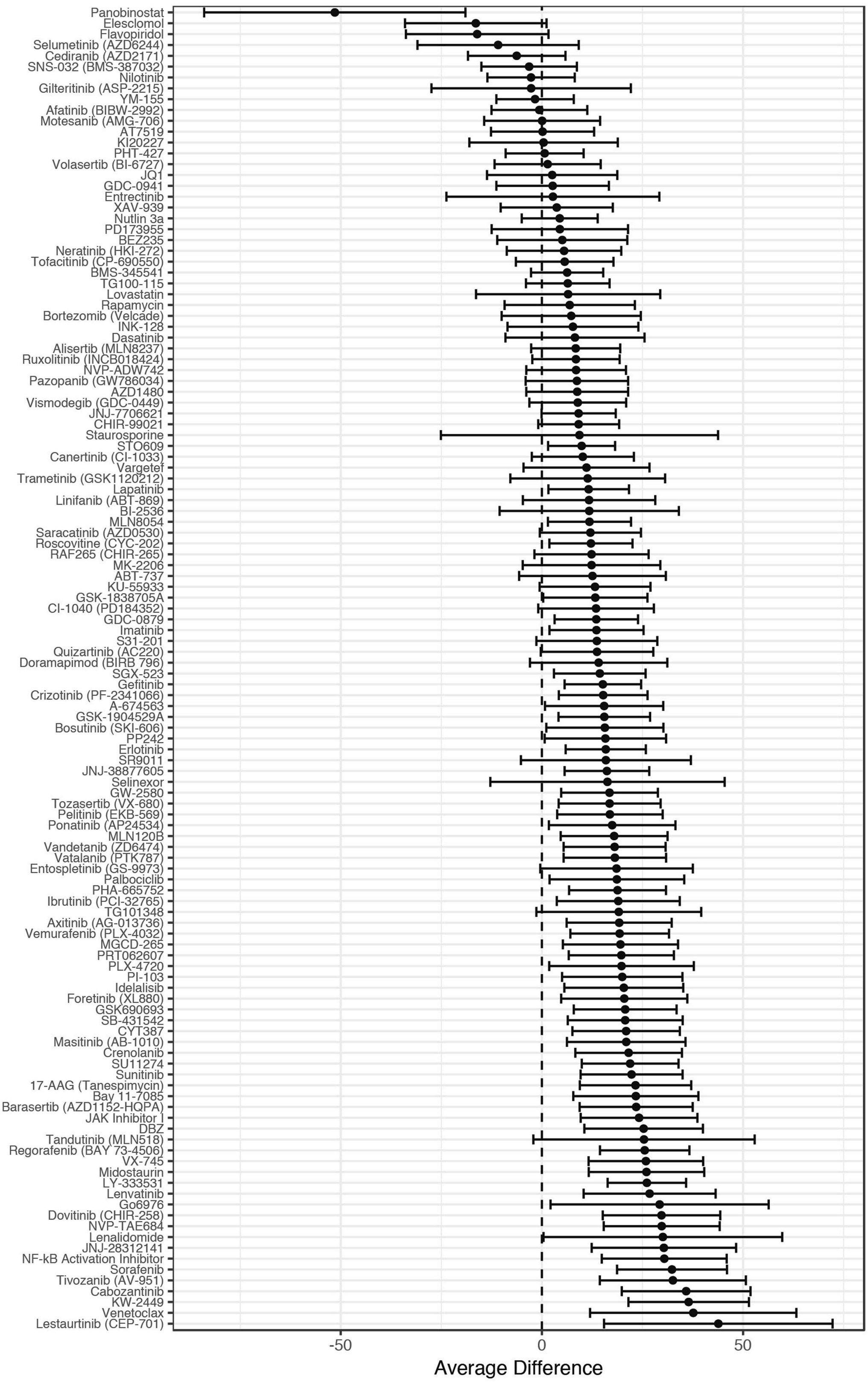
Extended Data Fig. 1 | Genomic landscape of the Beat AML cohort. In total, 622 specimens from 531 patients were used for whole-exome sequencing. Automated and manual curation steps (described in the Methods, Supplementary Information and at http://vizome.org/additional_figures_BeatAML.html) were used to obtain a final set of high-confidence variants (annotated in Supplementary Table 7) and the earliest sample for each individual patient was used in this analysis. Clinical cytogenetics and gene fusion calls from RNA sequencing were used to curate recurrent gene rearrangements (Supplementary Information). The mutational profile for each patient is shown for frequency-ranked mutational events (top) and frequency-ranked gene rearrangements (bottom). The mosaic plot is annotated with clinical features of each case, such as diagnosis or relapse and de novo or transformed disease states, and the first bar chart on the right summarizes the cohort frequencies of mutational and gene rearrangement events. The last bar chart on the right summarizes the frequency of significant drug–mutation associations for the given gene across the cohort with drug sensitivity displayed in red and drug resistance displayed in blue. Eleven genes that have not previously been reported to be somatically mutated in cancer were observed with mutations at approximately 1% cohort frequency: CUB and Sushi multiple domains 2 (*CSMD2*), NAC alpha domain containing (*NACAD*), teneurin transmembrane protein 2 (*TENM2*), aggrecan (*ACAN*), ADAM metallopeptidase with thrombospondin type 1 motif 7 (*ADAMTS7*), immunoglobulin-like and fibronectin type III domain containing 1 (*IGFN1*), neurobeachin-like 2 (*NBEAL2*), poly(U) binding splicing factor 60 (*PUF60*), zinc-finger protein 687 (*ZNF687*), cadherin EGF LAG seven-pass G-type receptor 2 (*CELSR2*) and glutamate ionotropic receptor NMDA type subunit 2B (*GRIN2B*). For the number of samples used to correlate each drug with mutations, see Supplementary Table 17.



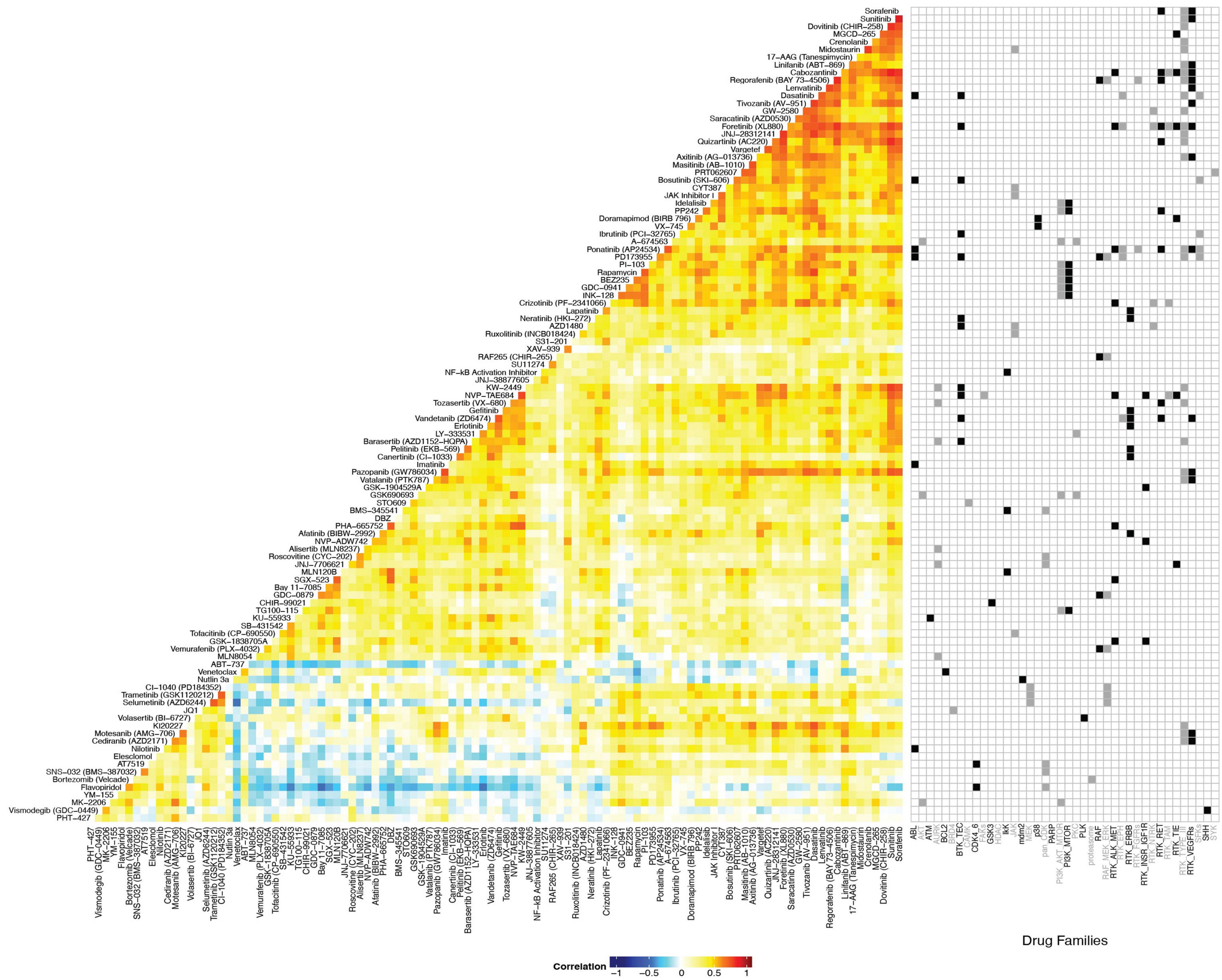
Extended Data Fig. 2 | Transcriptomic landscape of the Beat AML cohort. In total, 451 specimens from 411 patients with AML were used for RNA-sequencing analyses. The 2,000 genes with the greatest differential expression across these patients with AML are displayed as a heat map. The heat map is annotated with disease type, ELN risk stratification groups, and genetic and cytogenetic features of disease as indicated in the key.



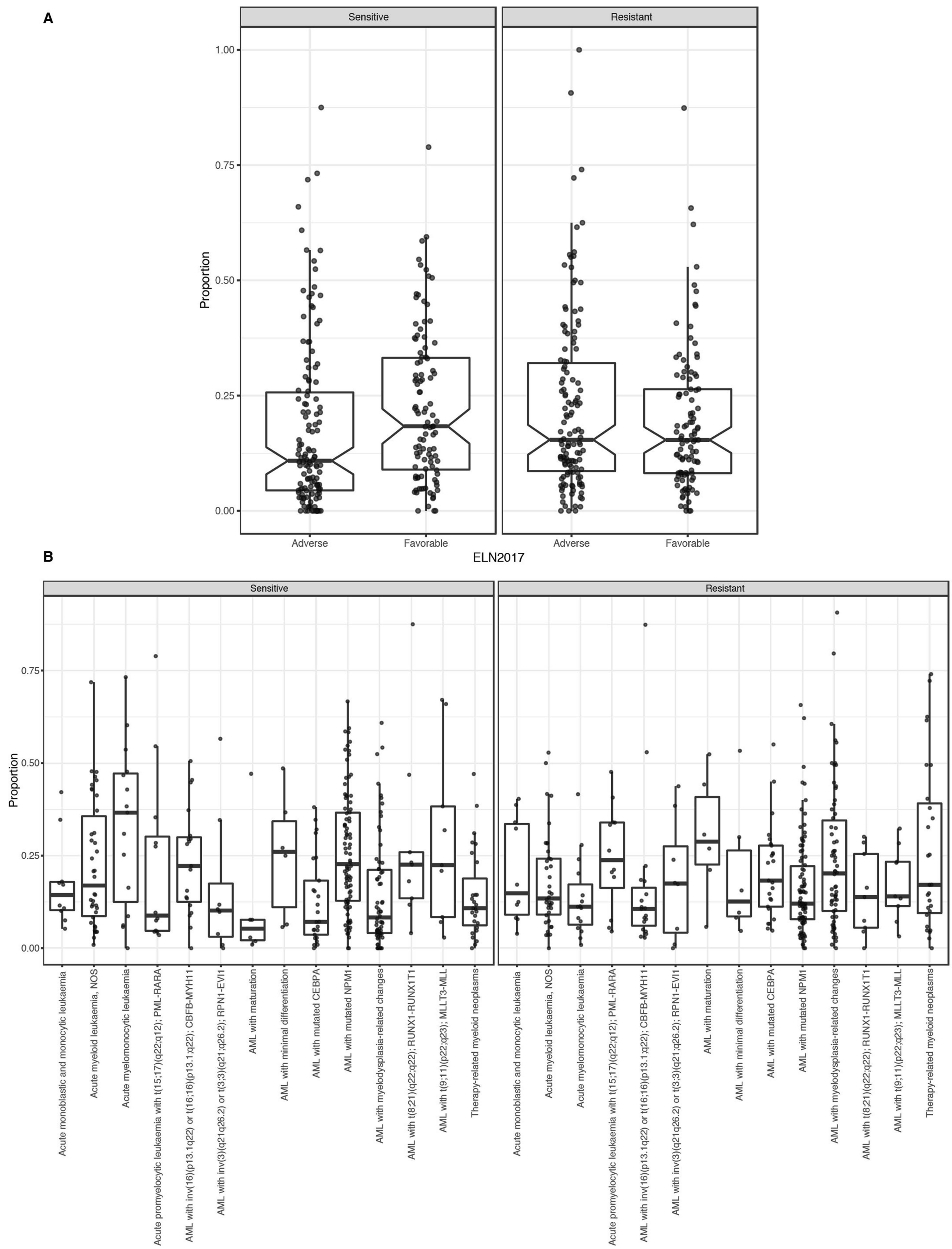
Extended Data Fig. 3 | Functional drug sensitivity landscape of the Beat AML cohort. In total, 409 specimens from 363 patients with AML were subjected to an ex vivo drug sensitivity assay, in which freshly isolated mononuclear cells from blood or bone marrow of patient specimens were incubated with graded concentrations of 122 small-molecule inhibitors (seven dose points in addition to the no drug control). Probit curve fits were used to compute drug-response metrics, and the *z* score of area under the dose–response curve is plotted for each individual patient specimen against each drug. Drug sensitivity (blue) and resistance (red) are annotated by a colour gradient, with grey indicating no drug data available. The heat map is annotated at the top and bottom with major clinical, cytogenetic and genetic features of disease as indicated in the key.



Extended Data Fig. 4 | Drug response in de novo versus transformed AML cases. The average inhibitor response AUCs for all cases that were de novo ($n = 288$) versus all cases that transformed from a background of myelodysplastic syndromes ($n = 111$) were compared for every inhibitor that had at least three cases with evaluable data in each group. The middle point represents the average difference in AUC between the two groups with the bars representing the 95% confidence interval. For the sample size and statistical results of each drug–sample group correlation, see Supplementary Table 20.

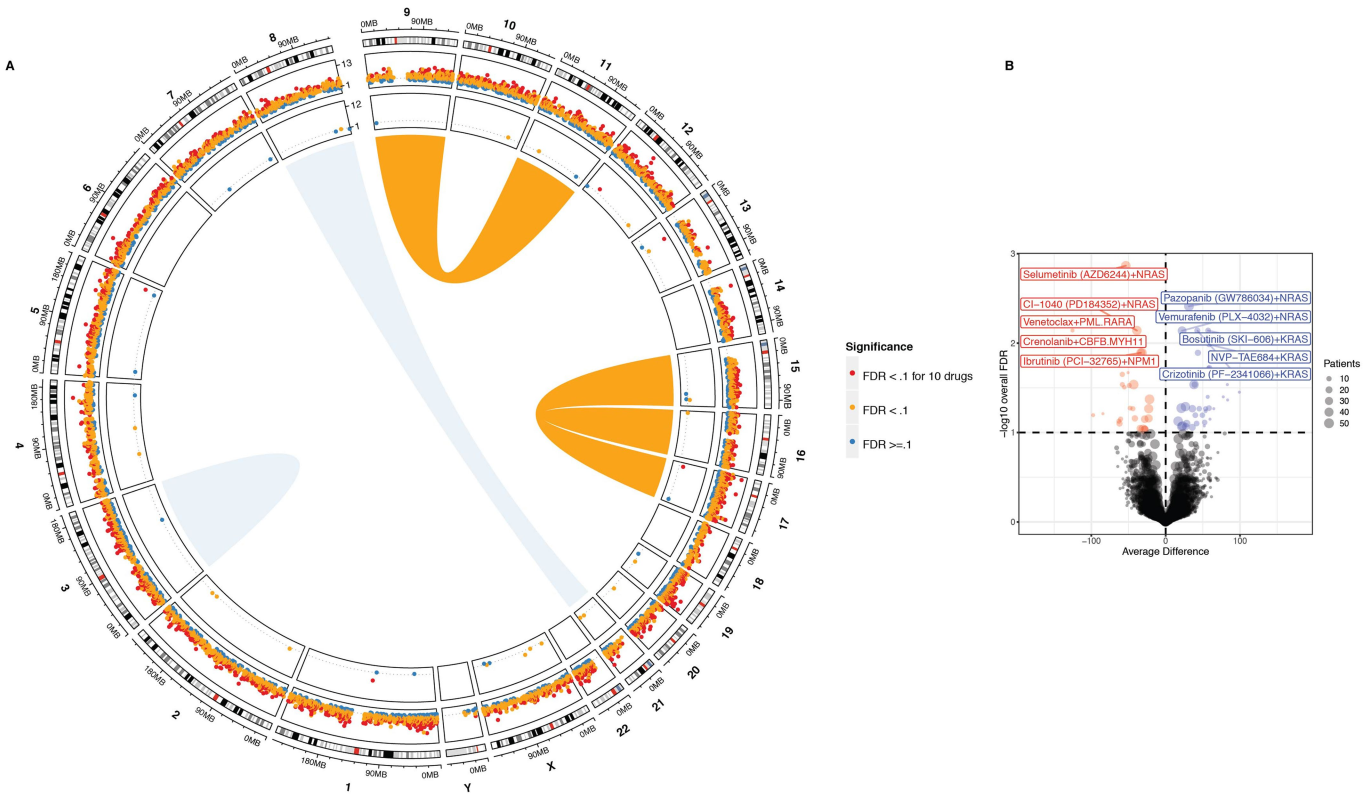


Extended Data Fig. 5 | Pairwise drug sensitivity correlations and association with drug family. To understand patterns of small-molecule sensitivity against prior annotations of the gene and pathway targets of each drug, drugs were placed into drug families according to target genes and/or pathways and the Pearson's correlation value of each drug was plotted onto a clustered heat map, showing drugs with similar or dissimilar patterns of sensitivity across the patient cohort. Annotations based on prior knowledge of the drug families to which each drug could be assigned are shown to the right of the heat map with alternating black and grey boxes and labels used to aid in tracking. Descriptions of each drug family as well as the number of samples used to calculate each pairwise drug correlation are found in Supplementary Tables 11, 21.

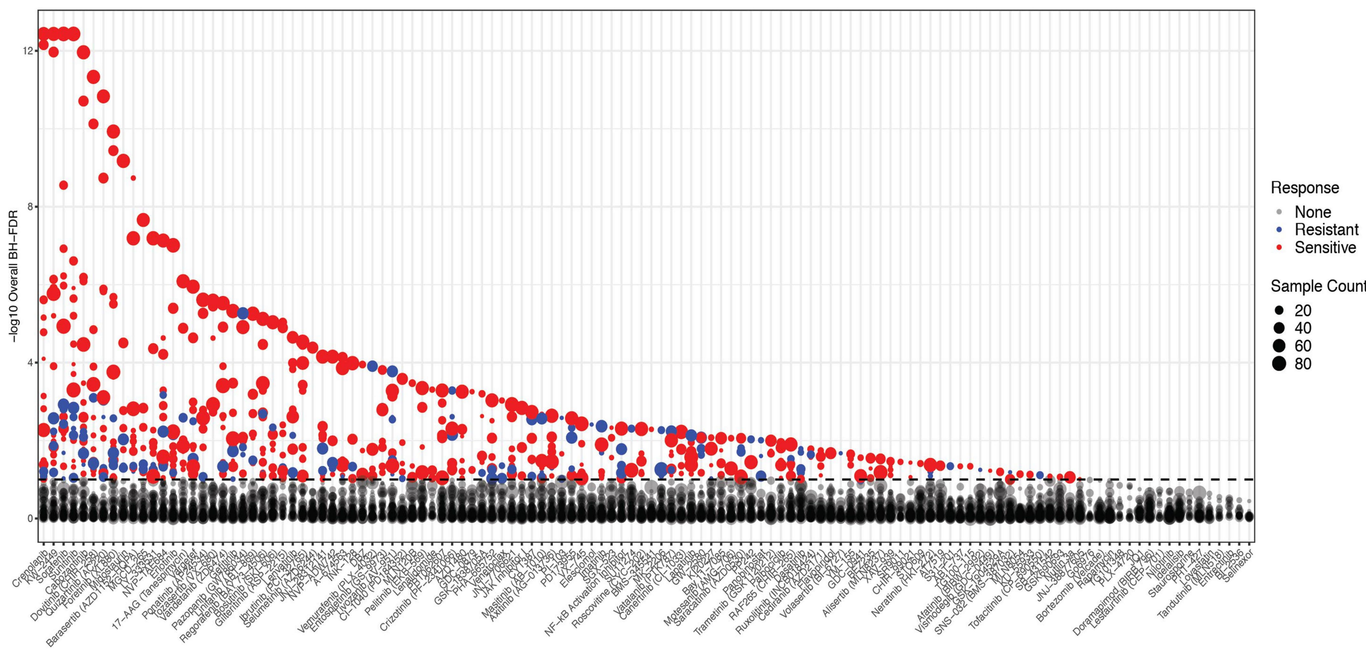


Extended Data Fig. 6 | See next page for caption.

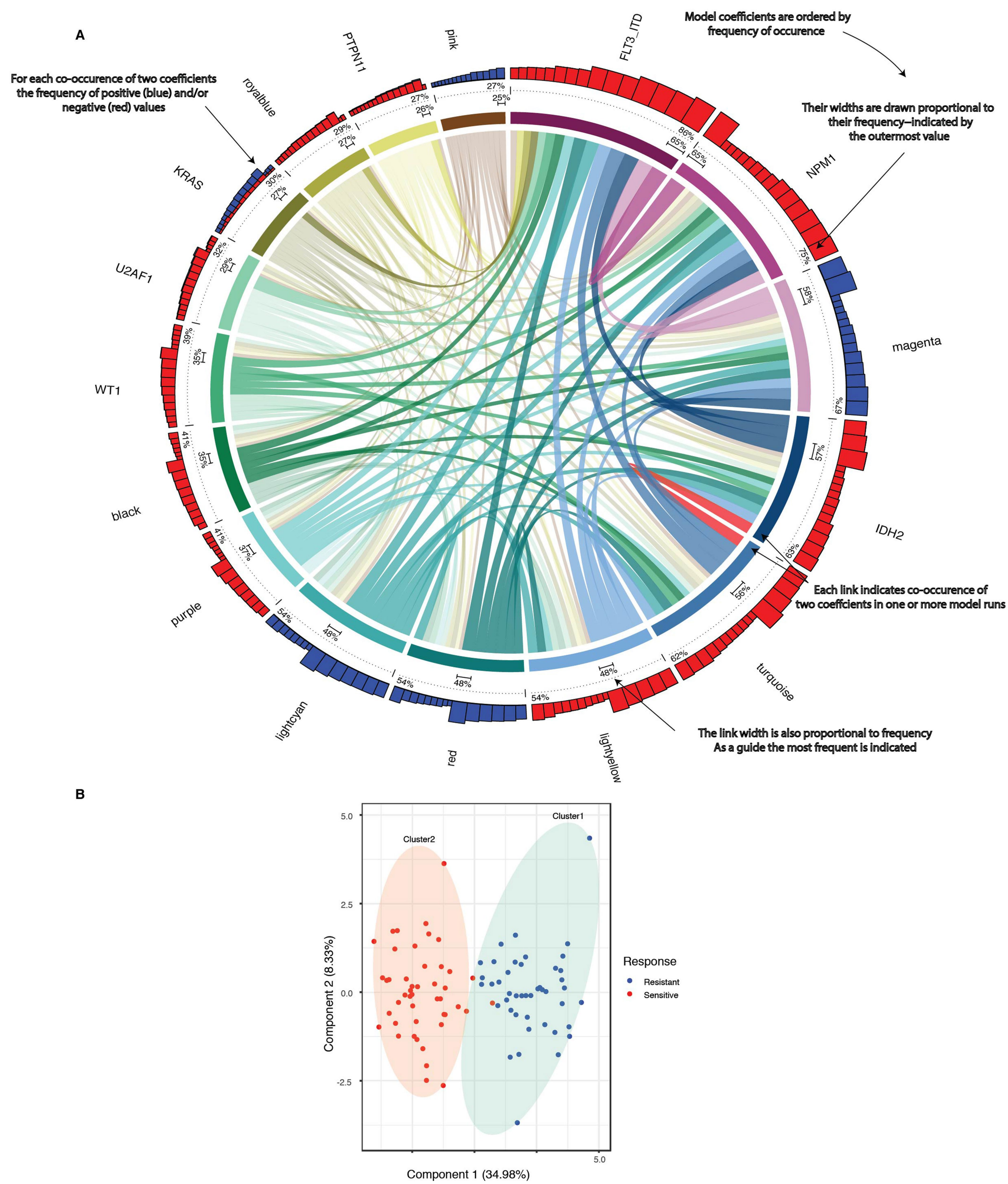
Extended Data Fig. 6 | Binary drug response calls and correlation with clinical subsets. a, For the intersect of every specimen with evaluable response data for each inhibitor, we created a threshold for binary sensitive or resistant calls based on whether the individual specimen response fell within the most sensitive 20% of all specimens tested against that drug. A matrix plot showing the unsupervised clustering of the binary calls can be found at http://vizome.org/additional_figures_BeatAML.html. The binary drug-resistance calls for each specimen were combined into a single value, representing the proportion of drugs to which an individual specimen was sensitive (left) or resistant (right). Specimens were divided into 'Favourable' and 'Adverse' groups based on ELN 2017 criteria to determine whether overall drug sensitivity or resistance correlated with prognostic features of disease ($n = 233$ patients). **b**, The binary drug-resistance calls for each specimen as in **a**. Specimens were divided into diagnostic groups based on WHO 2016 categories to determine whether overall drug sensitivity or resistance correlated with cytogenetic or morphologic features of disease ($n = 340$ patients). **a**, **b**, The top and bottom points of the box plots show 1.5 times the interquartile range (IQR) from the upper and lower lines; the top, middle and bottom lines indicate the 75th, median and 25th percentile of the data with the notches extending $1.58 \times \mathrm{IQR}/(\sqrt{(n)})$. Specific sample sizes of each group are reported in Supplementary Table 22.



Extended Data Fig. 7 | Integration of genetic events with drug sensitivity. a, Circos plot showing AML rearrangements in the centre, mutational events in the next concentric ring, and gene expression events in the outer ring. The size and width indicates frequency of the event and the FDR-corrected *Q* value of association with drug sensitivity is colour-coded (sensitivity (red); resistance (blue)). For each gene, tests involving expression were two-sided Student's *t*-tests (linear model) of the differences between sensitive and resistant samples. For mutational events, the average difference in AUC between mutant and wild-type samples was determined using two-sided Student's *t*-tests from a linear model as shown in Fig. 2a. FDR was computed using the Benjamini–Hochberg method over all the drugs. The number of samples used to correlate each mutational event with drug sensitivity is reported in Supplementary Table 17. **b**, As in Fig. 2a, the average difference in AUC drug response between mutant and wild-type cases was determined using a two-sided Student's *t*-test from a linear model fit (plotted on the *x* axis and the FDR-corrected *Q* value is plotted on the *y* axis). This analysis shows only FLT3-ITD-negative cases. FDR was computed using the Benjamini–Hochberg method over all the drugs. The number of samples used to correlate each mutational event with drug sensitivity is reported in Supplementary Table 17. Expanded and interactive plots are available in our online data browser (http://www.vizome.org/ and http://vizome.org/additional_figures_BeatAML.html).



Extended Data Fig. 8 | Integration of drug sensitivity with genetic events. Correlation between drug sensitivity and mutational events. The average difference in AUC drug response between mutant and wild-type cases was determined using a two-sided Student's *t*-test from a linear model fit. FDR was computed using the Benjamini–Hochberg method over all the drugs. The degree of significance is represented on the *y* axis (sensitivity (red); resistance (blue)). The number of samples used to correlate each mutational event with drug sensitivity is reported in Supplementary Table 17.



Extended Data Fig. 9 | Functional drug sensitivity landscape of the Beat AML cohort. a, Co-occurrences with regard to WGCNA gene expression clusters and/or mutational events (coefficients) were detected by multivariate modelling with respect to ibrutinib response (resistance (blue); sensitivity (red)) and the degree of correlation is shown in the stacked bar plot (top). All coefficients that appear in 25% of the bootstrapped sample sets are shown as segments of the circle. Segment width (the coloured ring) corresponds to the percentage of bootstrapped samples in which that coefficient appears (quantified above the dotted line). The variables appear in descending order clockwise starting at 12 o'clock. Each link indicates pairwise co-occurrence of mutational events and gene expression clusters (width represents frequency of the co-occurrence). The largest co-occurrence for each coefficient is quantified. **b,** The capacity of differential gene expression to distinguish the 20% most ibrutinib-sensitive ($n = 46$) from 20% most resistant ($n = 44$) specimens is shown on a principal component plot ($n = 239$ patient samples were tested for ibrutinib sensitivity and RNA sequencing). For the number of samples used to correlate each drug with gene expression and perform LASSO regression, see Supplementary Table 17.

ARTICLE

https://doi.org/10.1038/s41586-018-0616-y

Pathogen elimination by probiotic *Bacillus* via signalling interference

Pipat Piewngam[1,2], Yue Zheng[1,5], Thuan H. Nguyen[1,5], Seth W. Dickey[1], Hwang-Soo Joo[1,4], Amer E. Villaruz[1], Kyle A. Glose[1], Emilie L. Fisher[1], Rachelle L. Hunt[1], Barry Li[1], Janice Chiou[1], Sujiraphong Pharkjaksu[2], Sunisa Khongthong[3], Gordon Y. C. Cheung[1], Pattarachai Kiratisin[2] & Michael Otto[1]*

Probiotic nutrition is frequently claimed to improve human health. In particular, live probiotic bacteria obtained with food are thought to reduce intestinal colonization by pathogens, and thus to reduce susceptibility to infection. However, the mechanisms that underlie these effects remain poorly understood. Here we report that the consumption of probiotic *Bacillus* bacteria comprehensively abolished colonization by the dangerous pathogen *Staphylococcus aureus* in a rural Thai population. We show that a widespread class of *Bacillus* lipopeptides, the fengycins, eliminates *S. aureus* by inhibiting *S. aureus* quorum sensing—a process through which bacteria respond to their population density by altering gene regulation. Our study presents a detailed molecular mechanism that underlines the importance of probiotic nutrition in reducing infectious disease. We also provide evidence that supports the biological significance of probiotic bacterial interference in humans, and show that such interference can be achieved by blocking a pathogen's signalling system. Furthermore, our findings suggest a probiotic-based method for *S. aureus* decolonization and new ways to fight *S. aureus* infections.

There is increasing appreciation of the key role that the intestinal microbiota play in preventing the colonization and overgrowth of pathogens[1,2]. The mechanisms that have been implicated in this beneficial function of probiotic bacteria are mostly indirect, and include modulation of the immune system, enhancement of the intestinal epithelial barrier, or competition with pathogens for nutrients[2–5]. Whether there is direct interference between probiotic and pathogenic bacteria is less clear. Some probiotic strains produce bacteriocin proteins, which can kill phylogenetically related pathogenic bacteria[2], and it has been shown that a bacteriocin-producing *Escherichia coli* strain inhibits its colonization by related pathogenic bacteria in the inflamed gut of mice[6]. However, no evidence has been obtained to indicate that such mechanisms matter or are widespread in humans. Furthermore, it is not known whether there are mechanisms for direct probiotic bacterial interference that are not mediated by bacteriocins.

The genus *Bacillus* comprises different species of soil bacteria that form endospores with the ability to survive harsh environmental conditions, such as the high temperatures encountered during cooking procedures. *Bacillus* spores are commonly ingested with vegetables[7]. They can subsequently germinate to form metabolically active, vegetative cells[8], which can temporarily colonize the intestinal tract[9]. Given the variability in dietary customs, the concentration of *Bacillus* spores in human faeces is also highly variable. It has been reported to be around 10^5 colony-forming units (CFU) per gram on average, occasionally reaching up to 10^8 CFU per gram[7]. Several probiotic formulae contain *Bacillus* species[10], which are thought to reduce pathogen colonization by mechanisms that—except for a described immune-stimulatory effect on epithelial cells[11]—remain poorly defined.

Staphylococcus aureus is a widespread and dangerous human pathogen that can cause a variety of diseases, ranging from moderately severe skin infections to fatal pneumonia and sepsis[12]. Treatment of *S. aureus* infections is severely complicated by antibiotic resistance[13], such as in methicillin-resistant *S. aureus* (MRSA), and there is no working *S. aureus* vaccine[14]. Therefore, alternative strategies to combat *S. aureus* infections are eagerly sought[15]. Because *S. aureus* infections commonly originate from previous asymptomatic colonization[16,17], decolonization has recently gained considerable attention as a possible means to fight *S. aureus* infections in a preventive manner[18]. While the nares (nostrils) have traditionally been considered the primary *S. aureus* colonization site[19], there is increasing evidence that the intestinal tract is also commonly colonized by *S. aureus*[20–22] and forms an important reservoir for outbreaks of infectious *S. aureus* disease[23,24]. Several studies have reported levels of *S. aureus* in the faeces of human adults of around 10^3–10^4 CFU per gram[25–27]. Possibly, intestinal *S. aureus* colonization explains the failure of previous topical decolonization efforts aimed solely at the nose[16,22,28].

Here we hypothesized that the composition of the human gut microbiota affects intestinal colonization with *S. aureus*. To evaluate that hypothesis, we collected faecal samples from 200 healthy individuals from rural populations in Thailand (Fig. 1a). This exemplary population was selected in order to rule out, as much as possible, the food sterilization and antibiotic usage that are common in highly developed urban areas, which potentially could diminish the abundance of probiotic bacteria in the food and intestinal tracts of the participating subjects. Our analysis revealed a comprehensive *Bacillus*-mediated *S. aureus* exclusion effect in the human population. By demonstrating that quorum sensing is indispensable for *S. aureus* to colonize the intestine, and discovering that secreted *Bacillus* fengycin lipopeptides function as quorum-sensing blockers to achieve complete eradication of intestinal *S. aureus*, we provide evidence that strongly suggests that this pathogen-exclusion effect in humans is due to a widespread and efficient probiotic-mediated mechanism that inhibits pathogen quorum-sensing signalling.

S. aureus exclusion by Bacillus

We found that 25/200 (12.5%) of human subjects carried *S. aureus* in their intestines, as determined by growth from faecal samples. Nasal

[1]Pathogen Molecular Genetics Section, Laboratory of Bacteriology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. [2]Department of Microbiology, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand. [3]Faculty of Veterinary Science, Rajamangala University of Technology Srivijaya, Nakhon Si Thammarat, Thailand. [4]Present address: Department of Pre-PharmMed, College of Natural Sciences, Duksung Women's University, Seoul, South Korea. [5]These authors contributed equally: Yue Zheng, Thuan H. Nguyen. *e-mail: motto@niaid.nih.gov

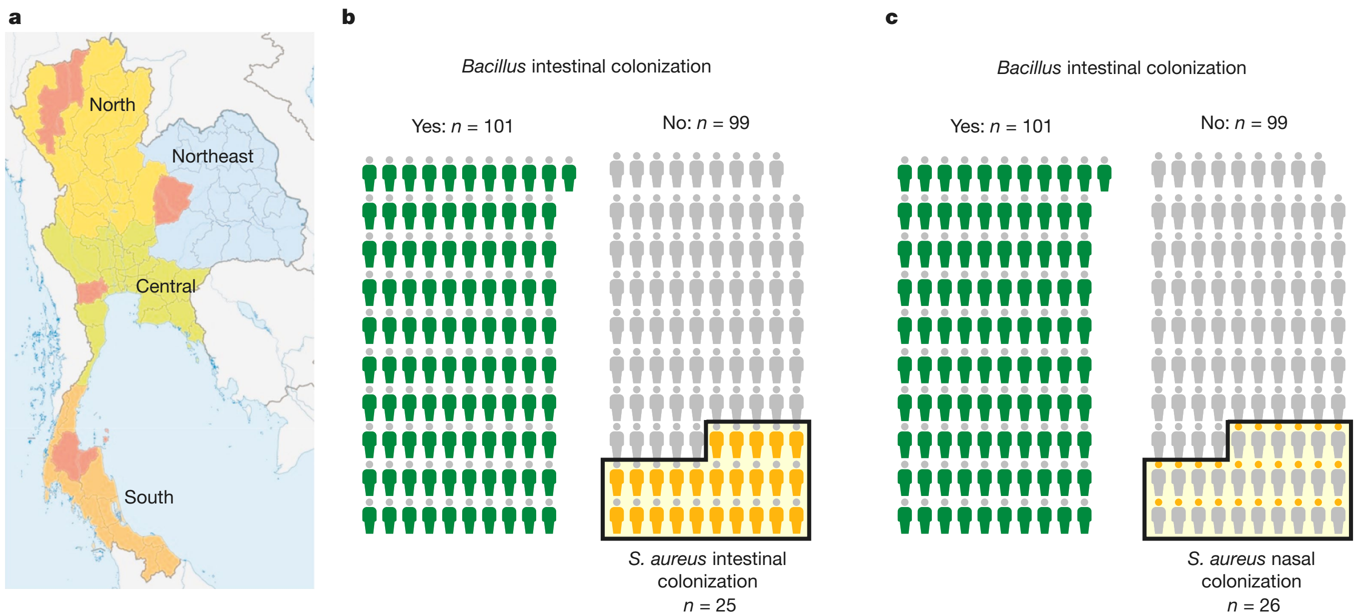


Fig. 1 | Exclusion of *S. aureus* colonization by dietary *Bacillus* in a human population. a, Areas (in red) from which faecal samples were collected in rural populations and analysed for the presence of *Bacillus* and *S. aureus*. **b**, **c**, Intestinal (**b**) and nasal (**c**) colonization with *S. aureus* (yellow) in individuals that showed (green) or did not show (grey) intestinal colonization with *Bacillus*.

carriage was similar in frequency (26/200; 13%), a result that is in accordance with previous findings showing a correlation between nasal and intestinal colonization[22]. These rates are considerably lower than those commonly found in adult populations during cross-sectional culture-based surveys that were performed mainly in hospital-admitted individuals in urbanized areas (on average, 20% for intestinal and 40% for nasal carriage)[16,21,22].

To examine the hypothesis that bacterial interactions in the gut determine intestinal *S. aureus* colonization, we first analysed the composition of the gut microbiome by 16S ribosomal RNA sequencing. However, we did not detect substantial differences in the composition of the microbiome between *S. aureus* carriers and non-carriers (Extended Data Fig. 1).

By contrast, we found a striking correlation between the presence of *Bacillus* bacteria and the absence of *S. aureus*. *Bacillus* species (mostly *B. subtilis*; Extended Data Table 1) were found in 101/200 (50.5%) of subject samples. *S. aureus* was never detected in faecal samples when *Bacillus* species were present ($P < 0.0001$, Fisher's exact test; Fig. 1b). Furthermore, this pathogen-exclusion effect was not limited to the site of interaction—the gut—but extended to *S. aureus* colonization in a general fashion. While *Bacillus* was generally absent from nasal samples, *S. aureus* nasal colonization was never detected when intestinal *Bacillus* was present ($P < 0.0001$, Fisher's exact test; Fig. 1c). Notably, the levels of *S. aureus* colonization that we found in non-*Bacillus*-colonized individuals from rural Thailand approximately match those reported—using similar culture-based assays—in urbanized Western areas. These findings indicate a widespread mechanism exerted by *Bacillus* species that comprehensively inhibits colonization with *S. aureus*. Moreover, they suggest that *S. aureus* colonization is increased in urban populations because of the lack of a probiotic, *Bacillus*-containing diet. Of particular note, the results also indicate that the intestinal site has a previously underappreciated role in determining general *S. aureus* colonization, a notion in accordance with findings attributing a key role to faecal transmission in MRSA recolonization[28].

When we analysed data from previous 16S rRNA-sequencing-based microbiome studies, we found strongly variant results and no correlation between the absence of *S. aureus* and the presence of *B. subtilis*: studies that reported considerable *B. subtilis* or *S. aureus* numbers (samples with more than 10% colonization by either species) did not reveal exclusion phenomena (average 14.89 ± 15.69% colonization by both species) (Extended Data Table 2). However, although we did not find a correlation, this might be due to the fact that such sequencing-based analyses are set up to detect high-order taxonomic shifts rather than specific differences on the species or genus level.

Quorum sensing and colonization

Our results, which show no substantial high-order taxonomic differences in the microbiome composition between *S. aureus* carriers and non-carriers, exclude an indirect effect of *Bacillus* on the microbiome composition. Rather, we hypothesized that the *Bacillus* isolates produce a substance that directly and specifically inhibits intestinal colonization by *S. aureus*. We first analysed whether there is a growth-inhibitory effect of the *Bacillus* isolates on *S. aureus*. However, only a minor growth inhibition occurred in just 6 out of 105 isolates (we saw a maximal 1-mm inhibition zone when using an agar diffusion test with a five-times-concentrated culture filtrate). Therefore, a growth-inhibitory effect fails to explain the observed complete correlation between the presence of *Bacillus* and the absence of *S. aureus*, and rules out a bacteriocin-mediated phenomenon.

The factors that are important for *S. aureus* intestinal colonization are poorly understood. One study in mice has implicated teichoic acids found in the bacterial cell wall, as well as the cell-surface protein clumping factor A (ClfA)[29]. Prompted by our previous finding that ClfA is positively regulated by the accessory gene regulator (Agr) quorum-sensing system[30], we hypothesized that the *Bacillus* isolates secrete a substance that interferes with quorum-sensing signalling. Quorum sensing is responsible for sensing the density of the bacterial population (the 'quorum') and controlling a concomitant alteration in cell physiology[31]. Because quorum-sensing signals and sensors differ between different types of bacteria[31], an underlying quorum-quenching mechanism could explain the specificity of the inhibitory effect that we detected.

Because the role of quorum sensing in *S. aureus* intestinal colonization is unknown, we first used a mouse model of *S. aureus* intestinal colonization to test whether Agr-based quorum sensing is involved (Fig. 2a). In all mouse models in our study, we included: first, a human faecal isolate belonging to a sequence type (ST) that was frequently detected in the faecal isolates that we obtained (ST2196),

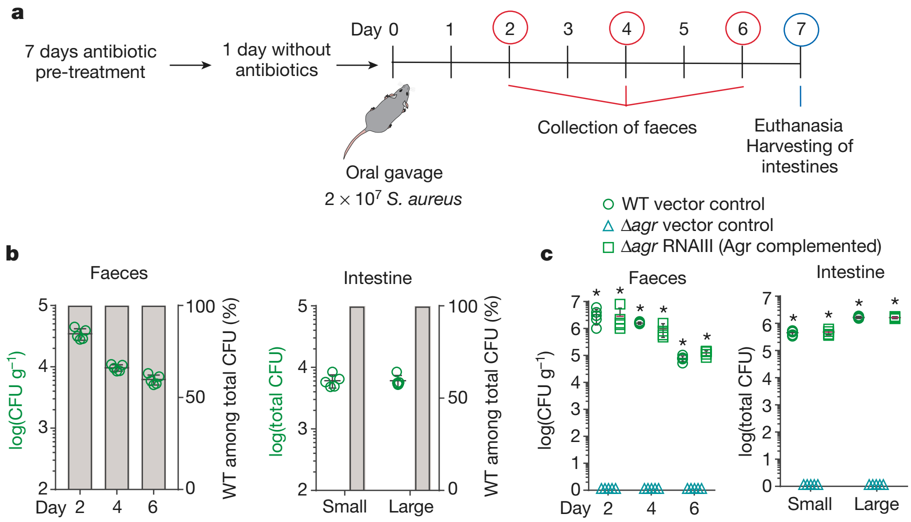


Fig. 2 | Quorum-sensing dependence of *S. aureus* intestinal colonization. a, Experimental set-up of the mouse intestinal colonization model. Mice received, by oral gavage, either 100 μl containing 10^8 CFU ml^{-1} of wild-type (WT) *S. aureus* strain ST2196 F12 and another 100 μl of 10^8 CFU ml^{-1} of the corresponding isogenic *agr* mutant ($n = 5$ per group; competitive experiment, shown in **b**); or 200 μl containing 10^8 CFU ml^{-1} wild-type, isogenic *agr* mutant or Agr (RNAIII)-complemented *agr* mutant ($n = 5$ per group; non-competitive experiment, shown in **c**). CFU in the faeces were determined two, four and six days after infection. At the end of the experiment (day seven), CFU in the small and large intestines were determined. **b**, Competitive experiment. Total obtained CFU are shown as dot plots; also shown are mean ± s.d. Bars show the percentage of wild-type among total determined CFU, of which 100 were analysed for tetracycline resistance (which is present only in the *agr* mutant). No *agr* mutants were detected in any experiment; therefore, all bars show 100% wild type. Given that 100 isolates were tested, the competitive index of wild type/*agr* mutant in all cases is ≥100. **c**, Non-competitive experiment with genetically complemented strains. Wild-type and isogenic *agr* mutant strains all harboured the pKX$_{\Delta}$16 control plasmid; Agr-complemented strains harboured pKX$_{\Delta}$RNAIII and constitutively expressed RNAIII, which is the intracellular effector of Agr. During the experiment, mice received 200 μg ml^{-1} kanamycin in their drinking water to maintain plasmids. Statistical analysis was performed using Poisson regression versus values obtained with the *agr* mutant strains. *$P < 0.0001$. Data are mean ± s.d. Note that no bacteria were found in the faeces or intestines of any mouse receiving *S. aureus* Δ*agr* with vector control. The corresponding zero values are plotted on the *x* axis of the logarithmic scale. See Extended Data Fig. 2 for corresponding data obtained using strains USA300 LAC and ST88 JSNZ.

according to multi-locus sequence typing (MLST) that we performed (Supplementary Table 1); second, a mouse infection isolate (ST88)[32]; and third, a human infection isolate of the highly virulent MRSA type USA300[33]. In competition experiments with equal amounts of wild-type and isogenic *agr* mutant strains, only wild-type *S. aureus* was detected in the faeces and colonized the large and small intestines at the end of the experiment (competition index ≥ 100) (Fig. 2b and Extended Data Fig. 2a, b). Furthermore, in a non-competitive experimental set-up, only those bacteria expressing the intracellular Agr effector RNAIII[34] achieved colonization; *agr*-negative control strains never did (Fig. 2c and Extended Data Fig. 2c). These data show that, in addition to its well-known role in infection[30,35], the Agr quorum-sensing system is absolutely indispensable for intestinal colonization.

Fengycin quorum quenchers

Having established that the Agr quorum-sensing regulatory system is essential for *S. aureus* intestinal colonization, we next analysed whether culture filtrates of the *Bacillus* isolates collected from human faeces can inhibit Agr. To that end, we used an *S. aureus* reporter strain, into the genome of which we had transferred the luminescence-conferring *luxABCDE* operon under the control of the Agr P3 promoter[34], which controls production of RNAIII. Remarkably, culture filtrates from all 105 isolates reduced Agr activity in the *S. aureus* reporter strain by at least 80% (Fig. 3a and Extended Data Table 1). No growth effects were observed, substantiating that growth inhibition does not underlie the inhibitory phenotype. Furthermore, a culture filtrate from a reference *B. subtilis* strain suppressed the production of key Agr-regulated virulence factors (phenol-soluble modulins, α-toxin and Panton–Valentine leucocidin; Fig. 3b, c and Supplementary Fig. 1). These results indicate that the inhibitory effect of the *Bacillus* isolates on *S. aureus* colonization is due to a secreted substance that inhibits Agr signalling.

To characterize the Agr-inhibitory substance(s), we performed experiments with culture filtrate of the reference *B. subtilis* strain. We found that the substance in question was thermostable and resistant to protease digestion (Extended Data Fig. 3a). In reversed-phase high-performance chromatography (RP-HPLC) (Extended Data Fig. 3b), substantial Agr-inhibiting activity was associated with two peaks, which we analysed by RP-HPLC/electrospray ionization mass spectrometry (ESI-MS) (Extended Data Fig. 3c). This analysis, together with the elution behaviour and published literature[36], allowed us to identify the Agr-inhibiting substances as members of the fengycin cyclic lipopeptide family. Because fengycins can differ in specific amino acids and in the length of the attached fatty acid, which usually is β-hydroxylated (β-OH), and because different *Bacillus* strains produce different fengycin species[37], we used further tandem mass spectrometric fragmentation analysis (MS/MS) to identify the specific fengycins present in the two active peaks (Extended Data Fig. 3d). Fengycins in the first peak were identified as β-OH-C17-fengycin A and β-OH-C16-fengycin B. The second peak consisted of one fengycin species, β-OH-C17-fengycin B. According to RP-HPLC/ESI-MS analysis, smaller, adjacent peaks also contained fengycin species, which we tentatively identified as β-OH-C17-fengycin A and the dehydroxylated versions of the identified three major fengycins (Extended Data Fig. 3e). For further analyses, we purified higher amounts of β-OH-C17-fengycin B to homogeneity from culture filtrate and verified the dose-dependent Agr-inhibiting activity of this pure substance (Extended Data Fig. 4).

Using RP-HPLC/ESI-MS analysis, we found fengycin production in all isolates, substantiating the general character of the inhibitory interaction (Extended Data Table 1). Although the production pattern of different fengycins varied between the analysed isolates, in many of them β-OH-C17-fengycin B was the most strongly produced type. Notably, almost complete inhibition of Agr was detected at a concentration of about 1.4 μM total fengycin (Fig. 3d). This corresponds to the median concentration of total fengycin (1.5 μM) produced by stationary-phase cultures of the *Bacillus* isolates (Fig. 3e).

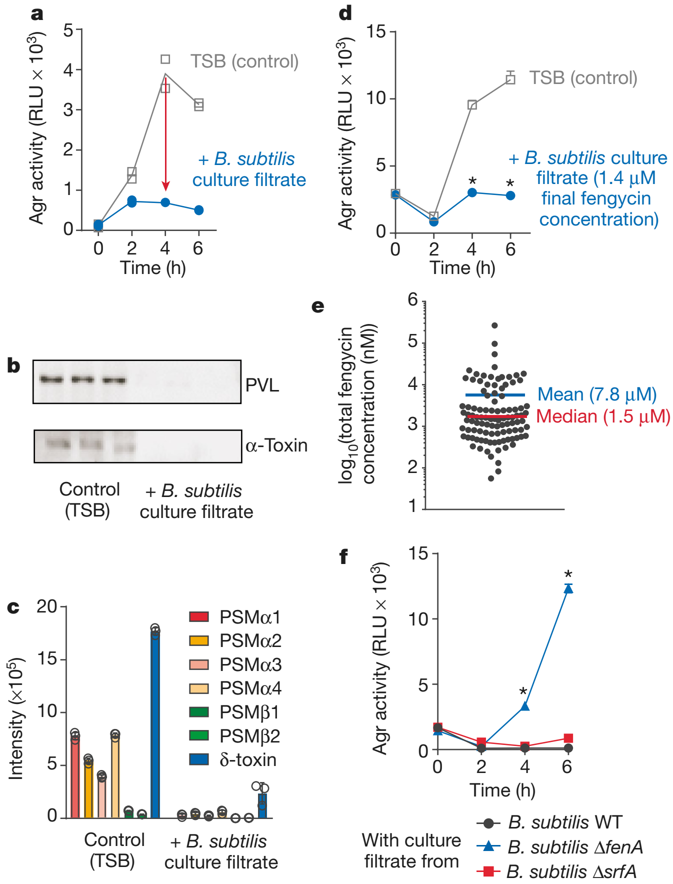


Fig. 3 | Inhibition of *S. aureus* quorum sensing by *Bacillus* fengycin lipopeptides. a, Example of an Agr-inhibition experiment. The *Bacillus* isolate was considered inhibitory if luminescence after 4-h growth of *S. aureus* was less than or equal to half of the control value (red arrow). RLU, relative light units; TSB, tryptic soy broth (control conditions). The experiment was performed with $n = 2$ biologically independent samples. The lines connect the means. **b**, Inhibition of expression of Panton–Valentine leucocidin (PVL) and α-toxin, using culture filtrate from the *B. subtilis* reference strain. Western blot analysis of $n = 3$ biologically independent samples was performed with filtrates from *S. aureus* cultures that had been grown for 4 h. See Supplementary Fig. 1 for the entire blots. **c**, Inhibition of expression of phenol-soluble modulins (PSMs) using culture filtrate from the *B. subtilis* standard strain. PSM expression was determined by RP-HPLC/ESI-MS after 4 h of *S. aureus* growth. **d**, Test for Agr-inhibitory capacity of *Bacillus* culture filtrate applied at a final concentration that represents the median concentration of total fengycin in the tested 106 *Bacillus* isolates. $^*P < 0.0001$ (two-way analysis of variance (ANOVA) with Tukey's post-test versus control). **e**, Total fengycin concentrations in stationary-phase culture filtrates of the 106 *Bacillus* isolates (see Extended Data Table 1 for details). **f**, Agr-inhibiting activities of *B. subtilis* wild-type (WT) in comparison to Δ*fenA* (fengycin-deficient) and Δ*srfA* (surfactin-deficient) strains. $^*P < 0.0001$ (two-way ANOVA with Tukey's post-test versus wild type). The experiments shown in **c**, **d**, **f** were performed with $n = 3$ biologically independent samples. Data are mean ± s.d.

To provide definitive evidence that fengycin production underlies the Agr-inhibiting capacity of *Bacillus*, we produced an isogenic mutant in the reference *B. subtilis* strain of the *fenA* gene, which is essential for fengycin production[38]. RP-HPLC/ESI-MS showed a specific absence of fengycins in that mutant strain, whereas surfactins—the predominant *Bacillus* lipopeptides—were still present (Extended Data Fig. 3f). Culture filtrate of the *fenA* mutant strain was devoid of Agr-inhibiting activity, in contrast to that of the isogenic wild-type strain (Fig. 3f). We also measured an isogenic surfactin-negative mutant strain, which showed Agr-inhibiting activity similar to that of the wild-type strain (Fig. 3f). These results confirmed that fengycin production is the source of the observed Agr inhibition.

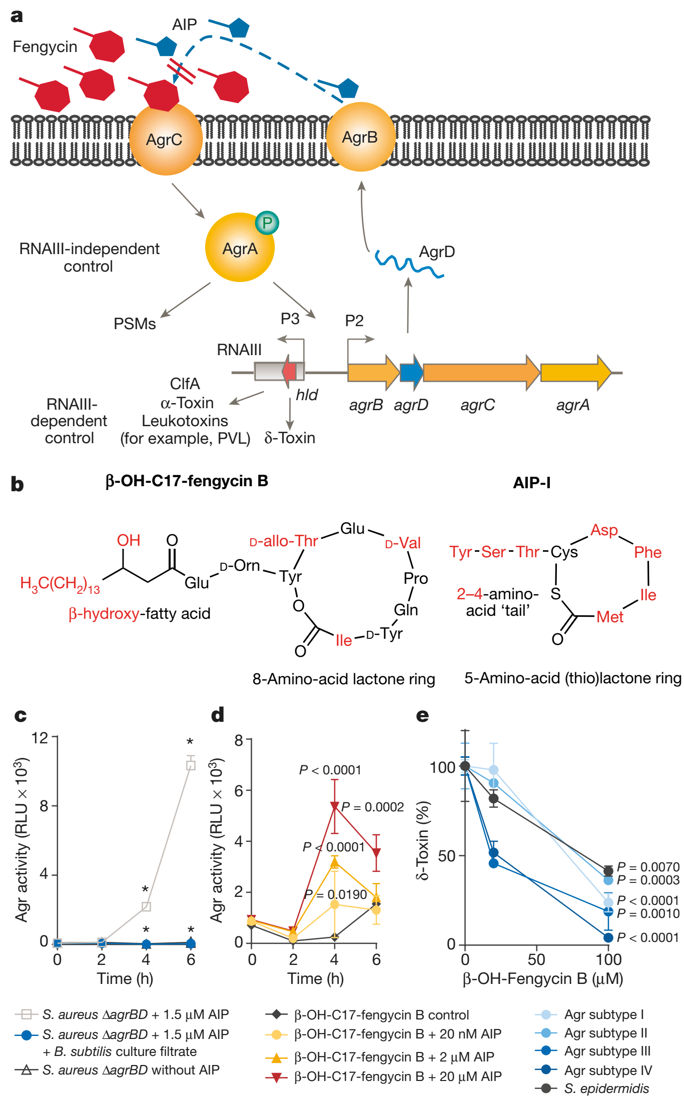


Fig. 4 | Competitive inhibition of *S. aureus* AIP activity by fengycins. a, Model of competitive Agr inhibition by fengycins. The *agrBDCA* operon (bottom right), whose expression is driven by the P2 promoter, encodes the AgrD precursor of the autoinducing-peptide (AIP), which is modified and secreted by AgrB. AIP binds to membrane-located AgrC, which, upon autophosphorylation, triggers phosphorylation and activation of the DNA-binding protein AgrA. In addition to stimulating transcription from the P2 promoter (autoinduction), AgrA drives expression of RNAIII, which in turn regulates the expression of target genes such as those encoding ClfA, α-toxin and leukotoxins. RNAIII also encodes the δ-toxin. Furthermore, AgrA drives the expression of phenol-soluble modulins (PSMs) in an RNAIII-independent fashion. **b**, Structural similarity of fengycins with AIPs. The structures of β-OH-C17-fengycin B and AIP-I are shown as examples. In red are structures and/or amino acids that may differ in different subtypes. **c**, Fengycins work by inhibiting AgrC. Shown is the inhibition of Agr by fengycin-containing *Bacillus* culture filtrate, using an *agrBD*-deleted *S. aureus* strain in which AgrC was stimulated by exogenously added AIP. $^*P < 0.0001$ (two-way ANOVA with Tukey's post-test; values obtained in Δ*agrBD*/AIP versus Δ*agrBD*/control (no AIP), and Δ*agrBD*/AIP/culture filtrate versus Δ*agrBD*/AIP). **d**, Competitive titration of fengycin-mediated Agr inhibition by increasing amounts of AIP, as assayed by the Agr luminescence assay. RLU, relative light units. Statistical analysis is by two-way ANOVA with Tukey's post-test versus control. **e**, Inhibition of Agr in different Agr-subtype *S. aureus* and *S. epidermidis* (strain 1457) by β-OH-C17-fengycin B, as measured by relative expression of δ-toxin using RP-HPLC/ESI-MS. Statistical analysis is by two-way ANOVA with Tukey's post-test versus intensity values obtained without addition of fengycin. Values were calculated as percentages relative to intensity values obtained without addition of fengycin, owing to different δ-toxin expression levels in the different strains. **c–e**, Experiments were performed with $n = 3$ biologically independent samples. Data are mean ± s.d.

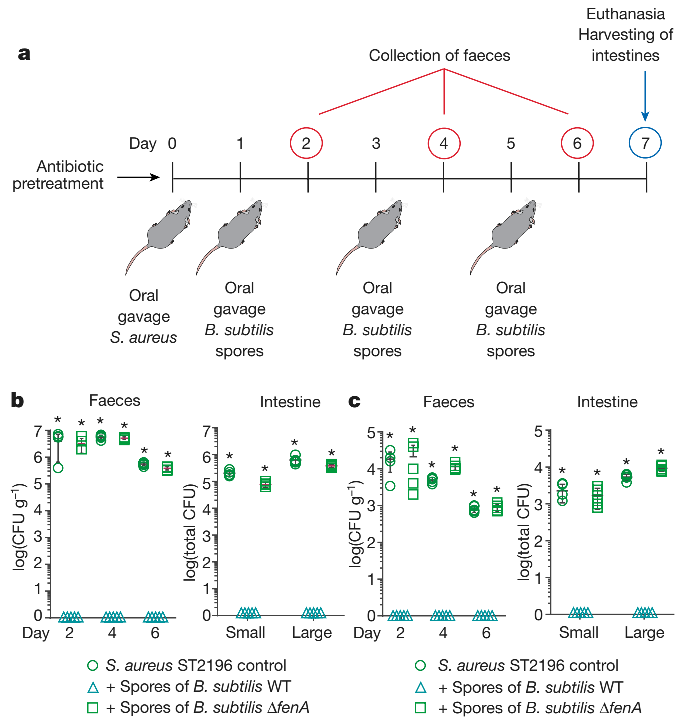


Fig. 5 | Inhibition of *S. aureus* colonization by dietary fengycin-producing *Bacillus* spores in a mouse model. a, Experimental set-up. $n = 5$ mice per group received 200 μl of 10^8 CFU ml^{-1} *S. aureus* strain ST2196 F12 by oral gavage. On the next day and every following second day, they received 200 μl of 10^8 CFU ml^{-1} spores of the *B. subtilis* wild-type (WT) or its isogenic *fenA* mutant, also by oral gavage. CFU in the faeces were determined two, four and six days after infection. At the end of the experiment (day seven), CFU in the small and large intestines were determined. The experiment was performed with (**b**) or without (**c**) antibiotic pretreatment. **b**, **c**, Experimental results. Statistical analysis was performed using Poisson regression versus values obtained with the *B. subtilis* WT spore samples. *$P < 0.0001$. Data are mean ± s.d. Note that no *S. aureus* were found in the faeces or intestines of any mouse challenged with *S. aureus* and receiving *Bacillus* wild-type spores. The corresponding zero values are plotted on the *x* axis of the logarithmic scale. See Extended Data Fig. 5 for corresponding data obtained using strains USA300 LAC and ST88 JSNZ.

Mechanism of fengycin-mediated inhibition

In the *S. aureus* Agr quorum-sensing regulatory circuit, the secreted Agr autoinducing peptide (AIP) interacts with an extracellular domain of AgrC, the histidine kinase part of a two-component signal-transduction system, to signal the cell-density status[39] (Fig. 4a). Different Agr subgroups of *S. aureus*, as well as different staphylococcal species, produce distinct cyclic heptapeptide to nonapeptide AIPs[35]. AIPs from other subgroups or species frequently inhibit Agr signal transduction by competitive inhibition at the AgrC-binding site[39–41]. Given that fengycins, being cyclic lipopeptides, show structural similarity to AIPs (Fig. 4b), it appears likely that fengycins compete with the natural AIP for AgrC binding. The only other theoretically possible site of interference from the extracellular space would be the membrane-located AIP production/secretion enzyme AgrB. Using an *S. aureus agrBD* deletion strain and stimulation of AgrC by synthetic AIP, which led to complete Agr activation, we ruled out that the target of Agr inhibition by *Bacillus* is AgrB (Fig. 4c). In further support of a mechanism that works through competition with AIP for binding to the AgrC receptor, we found that fengycin inhibition could be reversed in a dose-dependent fashion by adding AIP (Fig. 4d). Finally, we determined the AIP concentration in early stationary growth phase (at 6–8 hours) to be about 1 μM (Extended Data Fig. 5a), which is approximately equal to the concentration of fengycin for which we found complete Agr inhibition (Fig. 3d). These findings indicate that fengycins inhibit Agr signal transduction by efficient competitive inhibition as structural analogues of AIPs.

The fact that AgrC–AIP interaction differs according to Agr subtype raises the question of whether fengycins have a general ability to inhibit Agr. We found that purified β-OH-C-17 fengycin B inhibited Agr in members of all *S. aureus* Agr subtypes, as well as in *S. epidermidis* (Fig. 4e). Furthermore, the *S. aureus* strains used in our mouse experiments belong to different Agr subtypes (strain USA300, type I; strain ST88, type III; strain ST2196, type I). These results indicate that fengycins have broad-spectrum Agr-inhibiting activity.

Bacillus spores eradicate *S. aureus*

To validate our findings in vivo and demonstrate the specific role of fengycins in the inhibition of *S. aureus* intestinal colonization, we compared the impact of the *B. subtilis* wild-type reference strain and its isogenic *fenA* mutant on *S. aureus* colonization in a mouse intestinal colonization model. We first performed a control experiment to analyse the colonization kinetics of *B. subtilis* when given as spores, which corresponds to the form in which *Bacillus* would be taken up with food or probiotic formulae (Extended Data Fig. 5b). We observed transient colonization that strongly declined within two days. Importantly, colonization by the *B. subtilis fenA* mutant was not different to that by the wild-type strain, ruling out the possibility that fengycin production as such affects *B. subtilis* colonization.

Feeding mice *B. subtilis* spores completely abrogated colonization of all tested *S. aureus* strains in the faeces and intestines, in experimental set-ups with or without antibiotic pretreatment to eliminate the pre-existing microbiota. (Fig. 5b, c and Extended Data Fig. 5c–f). By contrast, spores of the *fenA* mutant had no notable effect on colonization of any *S. aureus* test strain. As *Bacillus* intestinal colonization in humans has been shown to reach much higher levels than that by *S. aureus*[7]—a situation likely to be even more pronounced in the tested rural population—our mouse data obtained with *S. aureus* numbers approximately equal to or exceeding those of applied *Bacillus* spores suggest that fengycin-mediated interference in quorum sensing contributes to the exclusion of *S. aureus* colonization that we observed in humans.

Conclusions

Scientific evidence to support the frequent claims that probiotic nutrients improve human health is scarce. However, this study provides evidence for a molecular mechanism by which probiotic bacteria found in food could directly interfere with pathogen colonization. In particular, our data underscore the often-debated[10,42] probiotic value of *B. subtilis*. Notably, we found the responsible agents to work by quorum quenching, demonstrating that pathogen exclusion in the gut may work by inhibition of a pathogen signalling system. Furthermore, our findings emphasize the importance of quorum sensing for pathogen colonization.

Our study suggests several valuable translational applications regarding alternative strategies to combat antibiotic-resistant *S. aureus*. First, the quorum-quenching fengycins—which previously had been known only for their antifungal activity[43]—could potentially be used as quorum-sensing blockers in eagerly sought antivirulence-based efforts to treat staphylococcal infections[15,44]. Second, *Bacillus*-containing probiotics could be used for simple and safe *S. aureus* decolonization strategies. In that regard, it is particularly noteworthy that our human data indicate that probiotic *Bacillus* can comprehensively eradicate intestinal as well as nasal *S. aureus* colonization. Such a probiotic approach would have numerous advantages over the present standard topical strategy involving antibiotics, which is aimed exclusively at decolonizing the nose[45].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0616-y.

Received: 11 June 2017; Accepted: 14 August 2018;
Published online 10 October 2018.

1. Guarner, F. & Malagelada, J. R. Gut flora in health and disease. *Lancet* **361**, 512–519 (2003).
2. Kamada, N., Chen, G. Y., Inohara, N. & Núñez, G. Control of pathogens and pathobionts by the gut microbiota. *Nat. Immunol*. **14**, 685–690 (2013).
3. Gourbeyre, P., Denery, S. & Bodinier, M. Probiotics, prebiotics, and synbiotics: impact on the gut immune system and allergic reactions. *J. Leukoc. Biol*. **89**, 685–695 (2011).
4. Macpherson, A. J. & Harris, N. L. Interactions between commensal intestinal bacteria and the immune system. *Nat. Rev. Immunol*. **4**, 478–485 (2004).
5. Bermudez-Brito, M., Plaza-Díaz, J., Muñoz-Quezada, S., Gómez-Llorente, C. & Gil, A. Probiotic mechanisms of action. *Ann. Nutr. Metab*. **61**, 160–174 (2012).
6. Sassone-Corsi, M. et al. Microcins mediate competition among *Enterobacteriaceae* in the inflamed gut. *Nature* **540**, 280–283 (2016).
7. Tam, N. K. et al. The intestinal life cycle of *Bacillus subtilis* and close relatives. *J. Bacteriol*. **188**, 2692–2700 (2006).
8. Casula, G. & Cutting, S. M. *Bacillus* probiotics: spore germination in the gastrointestinal tract. *Appl. Environ. Microbiol*. **68**, 2344–2352 (2002).
9. Duc, L. H., Hong, H. A., Barbosa, T. M., Henriques, A. O. & Cutting, S. M. Characterization of *Bacillus* probiotics available for human use. *Appl. Environ. Microbiol*. **70**, 2161–2171 (2004).
10. Hong, H. A., Duc, L. H. & Cutting, S. M. The use of bacterial spore formers as probiotics. *FEMS Microbiol. Rev*. **29**, 813–835 (2005).
11. Fujiya, M. et al. The *Bacillus subtilis* quorum-sensing molecule CSF contributes to intestinal homeostasis via OCTN2, a host cell membrane transporter. *Cell Host Microbe* **1**, 299–308 (2007).
12. Lowy, F. D. *Staphylococcus aureus* infections. *N. Engl. J. Med*. **339**, 520–532 (1998).
13. Lowy, F. D. Antimicrobial resistance: the example of *Staphylococcus aureus*. *J. Clin. Invest*. **111**, 1265–1273 (2003).
14. Septimus, E. J. & Schweizer, M. L. Decolonization in prevention of health care-associated infections. *Clin. Microbiol. Rev*. **29**, 201–222 (2016).
15. Dickey, S. W., Cheung, G. Y. C. & Otto, M. Different drugs for bad bugs: antivirulence strategies in the age of antibiotic resistance. *Nat. Rev. Drug Discov*. **16**, 457–471 (2017).
16. Wertheim, H. F. et al. The role of nasal carriage in *Staphylococcus aureus* infections. *Lancet Infect. Dis*. **5**, 751–762 (2005).
17. von Eiff, C., Becker, K., Machka, K., Stammer, H. & Peters, G. Nasal carriage as a source of *Staphylococcus aureus* bacteremia. *N. Engl. J. Med*. **344**, 11–16 (2001).
18. Simor, A. E. & Daneman, N. *Staphylococcus aureus* decolonization as a prevention strategy. *Infect. Dis. Clin. North Am*. **23**, 133–151 (2009).
19. Williams, R. E. Healthy carriage of *Staphylococcus aureus*: its prevalence and importance. *Bacteriol. Rev*. **27**, 56–71 (1963).
20. Mody, L., Kauffman, C. A., Donabedian, S., Zervos, M. & Bradley, S. F. Epidemiology of *Staphylococcus aureus* colonization in nursing home residents. *Clin. Infect. Dis*. **46**, 1368–1373 (2008).
21. Eveillard, M. et al. Evaluation of a strategy of screening multiple anatomical sites for methicillin-resistant *Staphylococcus aureus* at admission to a teaching hospital. *Infect. Control Hosp. Epidemiol*. **27**, 181–184 (2006).
22. Acton, D. S., Plat-Sinnige, M. J., van Wamel, W., de Groot, N. & van Belkum, A. Intestinal carriage of *Staphylococcus aureus*: how does its frequency compare with that of nasal carriage and what is its clinical impact? *Eur. J. Clin. Microbiol. Infect. Dis*. **28**, 115–127 (2009).
23. Senn, L. et al. The stealthy superbug: the role of asymptomatic enteric carriage in maintaining a long-term hospital outbreak of ST228 methicillin-resistant *Staphylococcus aureus*. *MBio* **7**, e02039-e15 (2016).
24. Squier, C. et al. *Staphylococcus aureus* rectal carriage and its association with infections in patients in a surgical intensive care unit and a liver transplant unit. *Infect. Control Hosp. Epidemiol*. **23**, 495–501 (2002).
25. Lindberg, E. et al. High rate of transfer of *Staphylococcus aureus* from parental skin to infant gut flora. *J. Clin. Microbiol*. **42**, 530–534 (2004).
26. Bhalla, A., Aron, D. C. & Donskey, C. J. *Staphylococcus aureus* intestinal colonization is associated with increased frequency of *S. aureus* on skin of hospitalized patients. *BMC Infect. Dis*. **7**, 105 (2007).
27. Ray, A. J., Pultz, N. J., Bhalla, A., Aron, D. C. & Donskey, C. J. Coexistence of vancomycin-resistant enterococci and *Staphylococcus aureus* in the intestinal tracts of hospitalized patients. *Clin. Infect. Dis*. **37**, 875–881 (2003).
28. Klotz, M., Zimmermann, S., Opper, S., Heeg, K. & Mutters, R. Possible risk for re-colonization with methicillin-resistant *Staphylococcus aureus* (MRSA) by faecal transmission. *Int. J. Hyg. Environ. Health* **208**, 401–405 (2005).
29. Misawa, Y. et al. *Staphylococcus aureus* colonization of the mouse gastrointestinal tract is modulated by wall teichoic acid, capsule, and surface proteins. *PLoS Pathog*. **11**, e1005061 (2015).
30. Cheung, G. Y., Wang, R., Khan, B. A., Sturdevant, D. E. & Otto, M. Role of the accessory gene regulator agr in community-associated methicillin-resistant *Staphylococcus aureus* pathogenesis. *Infect. Immun*. **79**, 1927–1935 (2011).
31. Miller, M. B. & Bassler, B. L. Quorum sensing in bacteria. *Annu. Rev. Microbiol*. **55**, 165–199 (2001).
32. Holtfreter, S. et al. Characterization of a mouse-adapted *Staphylococcus aureus* strain. *PLoS One* **8**, e71142 (2013).
33. Diep, B. A. et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant *Staphylococcus aureus*. *Lancet* **367**, 731–739 (2006).
34. Dastgheyb, S. S. et al. Role of phenol-soluble modulins in formation of *Staphylococcus aureus* biofilms in synovial fluid. *Infect. Immun*. **83**, 2966–2975 (2015).
35. Novick, R. P. & Geisinger, E. Quorum sensing in staphylococci. *Annu. Rev. Genet*. **42**, 541–564 (2008).
36. Pathak, K. V., Keharia, H., Gupta, K., Thakur, S. S. & Balaram, P. Lipopeptides from the banyan endophyte, *Bacillus subtilis* K1: mass spectrometric characterization of a library of fengycins. *J. Am. Soc. Mass Spectrom*. **23**, 1716–1728 (2012).
37. Cochrane, S. A. & Vederas, J. C. Lipopeptides from *Bacillus* and *Paenibacillus spp*.: a gold mine of antibiotic candidates. *Med. Res. Rev*. **36**, 4–31 (2016).
38. Chang, L. K. et al. Construction of Tn917ac1, a transposon useful for mutagenesis and cloning of *Bacillus subtilis* genes. *Gene* **150**, 129–134 (1994).
39. Lyon, G. J., Wright, J. S., Muir, T. W. & Novick, R. P. Key determinants of receptor activation in the agr autoinducing peptides of *Staphylococcus aureus*. *Biochemistry* **41**, 10095–10104 (2002).
40. Ji, G., Beavis, R. & Novick, R. P. Bacterial interference caused by autoinducing peptide variants. *Science* **276**, 2027–2030 (1997).
41. Otto, M., Echner, H., Voelter, W. & Götz, F. Pheromone cross-inhibition between *Staphylococcus aureus* and *Staphylococcus epidermidis*. *Infect. Immun*. **69**, 1957–1960 (2001).
42. Brisson, J. HSO's part 2—is Bacillus subtilis dangerous? *Fix Your Gut* http://fixyourgut.com/hso-probiotics-part-2-danger-supplementing-bacillus-subtilis/ (2014).
43. Vanittanakom, N., Loeffler, W., Koch, U. & Jung, G. Fengycin—a novel antifungal lipopeptide antibiotic produced by *Bacillus subtilis* F-29-3. *J. Antibiot*. **39**, 888–901 (1986).
44. Khan, B. A., Yeh, A. J., Cheung, G. Y. & Otto, M. Investigational therapies targeting quorum-sensing for the treatment of *Staphylococcus aureus* infections. *Expert Opin. Investig. Drugs* **24**, 689–704 (2015).
45. Poovelikunnel, T., Gethin, G. & Humphreys, H. Mupirocin resistance: clinical implications and potential alternatives for the eradication of MRSA. *J. Antimicrob. Chemother*. **70**, 2681–2692 (2015).

Acknowledgements We thank R. Kolter, Harvard Medical School, for providing the *B. subtilis srfA* mutant; D. Dubnau, Rutgers University, for the SPP1 phage; S. Holtfreter, University of Greifswald, and W. P. Zeng, Texas Tech University Health Sciences Center, for providing strain JSNZ/ST88; B. Krismer, University of Tübingen, for plasmid pKX15; F. DeLeo, National Institute of Allergy and Infectious Diseases (NIAID), for anti-Panton–Valentine-leucocidin; and N. A. Amissah for technical assistance. This work was supported by the Intramural Research Program of the NIAID, US National Institutes of Health (NIH) (project ZIA AI000904-16, to M.O.); and the Thailand Research Fund through the Royal Golden Jubilee PhD Program (grant number PHD/0072/2557, to P.P. and P.K.). P.K. was also supported by the Faculty of Medicine, Siriraj Hospital, Mahidol University, grant number (IO) R015833012; P.P. by the Graduate Partnership Program of the NIH; and S.W.D. by the Postdoctoral Research Associate Program of the National Institute of General Medical Sciences (1FI2GM11999101).

Reviewer information *Nature* thanks A. Baumler, M. Parsek and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions P.P., S.P. and S.K. collected human samples and analysed bacterial isolates by matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI–TOF MS). Y.Z., H.-S.J. and M.O. performed analytical and preparative chromatography. P.P., T.H.N., E.L.F., R.L.H., J.C. and G.Y.C.C. performed animal studies. S.W.D. constructed the *S. aureus agrBD* mutant and A.E.V. constructed all other *agr* mutants and complemented strains. K.A.G., A.E.V. and B.L. performed MLST. P.P. performed reporter assays, the microbiome study, and all further analyses not specifically mentioned. P.K. supervised the human analyses and M.O. all other parts of the study.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0616-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0616-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.O.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment, except for when noted.

Sample collection and bacterial screening. Nasal swabs and faecal samples were obtained from 200 Thai healthy volunteers from four different locations in southern, central, northeastern and northern Thailand. One sterile nasal swab, a sample collection tube, a sterile container and tissue paper were given to each participant. All participants provided informed written consent. The study was performed in compliance with all relevant ethical regulations and approved by the Siriraj Institutional Review Board (approval no. Si 733/2015). All participants were at least 20 years old (age range 20–87 years; median age 57 ± 14.5 years; 131 women and 69 men) and without history of intestinal disease. None had received any antibiotic treatment or stayed at a hospital within at least three months before the study.

Nasal swabs and faecal samples were streaked on mannitol salt agar (MSA) and then incubated at 37 °C for 24 h. Positive or negative *S. aureus* or *Bacillus* colonization could easily be distinguished by either strong growth on the entire plate, or the absence of any colonies, respectively. At the time of this analysis, the purpose was to obtain and archive colonizing *S. aureus* strains. As the hypothesis regarding *Bacillus*/*S. aureus* exclusion was developed only after we obtained the results of this analysis, the staff performing the analysis were blinded as to the exclusion hypothesis. Isolates were easily recognized as *S. aureus* or *Bacillus* by colony morphology and colour; however, every isolate was confirmed for species identity using matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI–TOF MS; see below), and *Bacillus* species were further distinguished by 16S rRNA sequencing (Extended Data Table 1). To that end, 16S rRNA genes were amplified by the polymerase chain reaction (PCR) using primers 27FB and 1492RAB[46] and similarity analysis with the basic local alignment search tool (BLAST) was used to identify the species. Subjects were considered as permanently colonized by *S. aureus* if two positive samples were obtained, tested after a four-week interval. All individuals tested either negative or positive for *S. aureus* at both times. In total, 105 *Bacillus* isolates from 101 individuals were analysed. In the samples from four individuals, two isolates each were taken owing to their apparent phenotypic differences.

Bacterial identification using MALDI–TOF MS. Isolates were inoculated onto sheep blood agar and incubated for 24 h at 37 °C. Bacterial colonies were applied onto a 96-spot target plate and allowed to dry at room temperature. Subsequently, 2 μl of MALDI matrix (a saturated solution of α-cyano-4-hydroxycinnamic acid (HCCA) in 50% acetonitrile and 2.5% trifluoroacetic acid) was applied onto the colonies and allowed to dry before testing. Then the target plate was loaded into the MALDI–TOF MS instrument (MicroFlex LT mass spectrometer, Bruker Daltonics). Spectra were analysed using MALDI Biotyper automation control and the Bruker Biotyper 2.0 software and library (version 2.0, 3,740 entries; Bruker Daltonics). Identification score criteria used were those recommended by the manufacturer: a score of $\geq$2.000 indicated species-level identification; a score of 1.700–1.999 indicated identification to the genus level; and a score of <1.700 was interpreted as no identification. Isolates that failed to produce a score of <1.700 with direct colony or extraction methods were retested. *S. aureus* ATCC25923, *E. coli* ATCC25922 and *Pseudomonas aeruginosa* ATCC27853 were used as controls.

Bacterial strains and growth conditions. The reference *B. subtilis* strain and parent of the *fenA* and *srfA* mutants used in this study was strain ZK3814 (genotype NCIB3610). The *S. aureus* strains used in all experiments (except the experiment in which we analysed different Agr-subtype *S. aureus*) were: first, the human faecal isolate F12 of ST2196 (Supplementary Table 1); second, strain JSNZ of ST88, a mouse isolate previously described as mouse adapted[32]; and third, strain LAC of pulsed-field type USA300, an MRSA lineage predominantly involved in community-associated infections, but now generally representing the major lineage responsible for *S. aureus* infections in the United States[47].

Isogenic mutants in *agr* were previously described (for strain LAC)[48] or produced in this study (for strains JSNZ and F12) by phage transduction of the *agr* deletion from strain RN6911. The *agr* system is entirely deleted in these strains, except for a 3′ part of RNAIII, which is not transcribed owing to the absence of the corresponding promoter. All mutants were verified by analytical PCR.

Owing to the tetracycline resistance introduced in the *agr* deletion strains, kanamycin derivatives (pKX$_{\Delta}$) of the pTX$_{\Delta}$ expression plasmid series were constructed and used for complementation of Agr. (This was not possible in strain LAC, which harbours resistance to multiple antibiotics, including kanamycin.) To that end, we treated plasmid pKX15[49]—provided by B. Krismer, University of Tübingen—as described[48] to delete the *xylR* repressor gene, in order to make expression of any fragment cloned under control of the *xyl* promoter constitutive. To obtain plasmid pKX$_{\Delta}$RNAIII, the RNAIII BamH1–MluI fragment was transferred from pTX$_{\Delta}$RNAIII[50]. Plasmid pKX$_{\Delta}$16 is the corresponding empty control plasmid, derived from pKX16 by analogous deletion of the *xylR* repressor gene.

To construct the *agrBD* deletion mutant of strain LAC P3-*lux*, we used a 4.8-kilobase PCR product from USA300 genomic DNA that included the *agrBDCA* operon as well as 1 kb upstream and 1 kb downstream; we cloned this product into the SmaI site of plasmid pIMAY[51] and used inverse PCR to delete *agrBD*. Allelic exchange was then performed, and the chromosomal deletion was confirmed by PCR using one primer outside of the 1-kb homology arm, followed by sequencing of the PCR product. See Supplementary Table 2 for the oligonucleotides used.

To construct the tetracycline-resistant derivatives of *S. aureus* ST88 and ST2196, we carried out φ11-phage-mediated transduction as described in order to transfer the tetracycline cassette in the donor strain (*S. aureus* RN4220 with integrated pLL29) to *S. aureus* strains ST88 and ST2196[52].

To construct the *B. subtilis* fengycin mutant strain, SPP1-phage-mediated transduction[53] was performed to transfer the *fenA* deletion present in the donor strain (BKE18340, a *fenA*(*ppsA*)::*erm* mutant in *B. subtilis* strain 168 obtained from the *Bacillus* Genetic Stock Center) to *B. subtilis* strain ZK3814. This was necessary as *B. subtilis* strain 168 bears a mutation in the *sfp* gene, abolishing lipopeptide production.

Bacteria were generally grown in tryptic soy broth (TSB) with shaking unless otherwise indicated.

Typing of *S. aureus* isolates. *S. aureus* isolates were typed by MLST as described[54]. PCR amplicons of seven *S. aureus* housekeeping genes (*arcC*, *aroE*, *glpF*, *gmk*, *pta*, *tpi* and *yqiL*) were obtained from chromosomal DNA and their sequences compared with those available from the PubMLST database (https://pubmlst.org/saureus/). Previously undescribed alleles (*arcC* 520–521 and *gmk* 337) and sequence types (ST4630–ST4638) were deposited to the website. The Agr subtype of *S. aureus* isolates was determined using a modified multiplex quantitative reverse transcription PCR (qRT–PCR) protocol[55]. Two duplex qRT–PCR protocols, using the respective described primer sets and two coloured probes each, were set up for Agr types I and II, and III and IV, respectively. Isolates for which the Agr type could not be determined by that method were analysed for the type of AIP production using RP-HPLC/ESI-MS with the chromatography method also used for PSM detection (see below), integrating the three major *m*/*z* peaks for each AIP type.

Microbiome analysis. Genomic DNA from each faecal sample was extracted using a QIAamp DNA stool Minikit (Qiagen) according to the manufacturer's instructions. The DNA was quantified using a Nanodrop spectrophotometer, and 16S rRNA paired-end sequencing of the V4 region of 16S rRNA was performed by Illumina using an Illumina MiSeq system as described[56].

For all obtained paired-end sequences, the abundance of operational taxonomic units (OTUs) and alpha and beta diversity were identified using quantitative insights into microbial ecology (QIIME 1.9.1)[57]. This study used the Nephele (release 1.6) platform from the National Institute of Allergy and Infectious Diseases (NIAID) Office of Cyber Infrastructure and Computational Biology (OCICB) in Bethesda, Maryland, USA. The sequences were assigned to OTUs with the QIIME's uclust-based[58] open-reference OUT-picking protocol[59] and the Greengenes 13_8 reference sequence set[60] at 99% similarity. Alpha diversity was calculated using Chao1 and Shannon analyses[61] and compared across groups using a non-parametric *t*-test with 999 permutations.

Growth-inhibition analysis. Growth inhibition of *S. aureus* by *Bacillus* culture filtrates was tested with an agar diffusion assay. To that end, 10 μl of *Bacillus* culture filtrate from each isolate was spotted on sterile filter disks. The filters were left to dry and the procedure was repeated four times, after which filters were laid on agar plates containing *S. aureus*, resulting in the analysis of five-times concentrated culture filtrate.

Fengycin purification. To identify the Agr-inhibiting active substance, 10 ml of culture filtrate from the *B. subtilis* reference strain grown for 48 h in TSB were applied to a Zorbax SB-C18 9.4 mm × 25 cm reversed-phase column (Agilent) using an AKTA Purifier 100 system (GE Healthcare). After washing with three column volumes of 100% buffer A (0.1% trifluoroacetic acid (TFA) in water) and five column volumes of 30% buffer B (0.1% TFA in acetonitrile), a 20-column volume gradient from 30% to 100% buffer B was applied. The column was run at a flow rate of 3 ml min^{-1}. Peak fractionation was performed using the absorbance at 214 nm, and fractions were subjected to further analysis by RP-HPLC/ESI-MS and MS/MS and tested for Agr inhibition (see below).

To purify larger amounts of the main active peak containing β-OH-C17-fengycin B, we added acetonitrile to 200 ml filtrate from cultures grown under the same conditions to a final concentration of 10%; precipitated material was removed by centrifugation for 10 min at 3,700*g* using a Sorvall Legend RT centrifuge, and the obtained cleared supernatant was applied to a self-packed HR 16/10 column filled with Resource PHE (GE Healthcare) material (column volume 17 ml). After sample application, the column was washed with 10% buffer B for three column volumes and 25% buffer B for five column volumes, after which a gradient of 15 column volumes from 25% to 60% buffer B was applied. We collected 10-ml fractions and lyophilized positive fractions (as determined by RP-HPLC/ESI-MS). The lyophilisate was redissolved in 2 ml acetonitrile. We added 6 ml of water and

removed the precipitated material through a 5-min centrifugation in a table-top centrifuge at maximum speed. The cleared supernatant was then further purified on a Zorbax SB-C18 9.4 mm × 25 cm reversed-phase column as described above.

PSM and lipopeptide detection by RP-HPLC/ESI-MS. PSMs were analysed by RP-HPLC/ESI-MS using an Agilent 1260 Infinity chromatography system coupled to a 6120 Quadrupole LC/MS in principle as described[62], but with a shorter column and a method that was adjusted accordingly. A 2.1 mm × 5 mm Perkin-Elmer SPP C8 (2.7 μm) guard column was used at a flow rate of 0.5 ml min^{-1}. After sample injection, the column was washed for 0.5 min with 90% buffer A and 10% buffer B, then for 3 min with 25% buffer B. Next, an elution gradient was applied from 25% to 100% buffer B in 2.5 min, after which the column was subjected to 2.5 min of 100% buffer B to finalize elution.

Bacillus culture filtrates or (partially) purified fractions containing lipopeptides (fengycins and surfactins) were analysed using the same column, system and elution conditions. To quantify the production of different fengycins, we used the two most abundant peaks, corresponding to double- and triple-charged ions, for the integration. Agilent mass hunter quantitative analysis version B.07.00 was used for quantification.

Measurement of Agr activity. To determine the Agr-inhibiting activity of *Bacillus* culture filtrates or purified fractions, we measured luminescence emitted by an Agr P3 promoter–*luxABCDE* reporter fusion construct that was inserted into the genome of *S. aureus* strain LAC[34]. Strain LAC P3–*luxABCDE* was diluted 100-fold from a preculture grown overnight in TSB before distribution into a 96-well microtitre plate. To 100 μl of that dilution, we added 100 μl of sterilized culture filtrate sample, unless otherwise indicated. Plates were incubated at 37 °C with shaking. Luminescence was measured with a GloMax Explorer luminometer (Promega) every 2 h for a total of 6 h. Inhibition was considered significant if the 4-h sample and control values differed by at least a factor of two. Of note, the quorum-quenching effect exerted by the one-time initial dose of fengycin or fengycin-containing culture filtrates was transient and was overcome at later times by the increasing intrinsic AIP production. The Agr-inhibiting activity of purified fengycin was also measured using quantitative real-time PCR of RNAIII as described[30].

To determine the Agr-inhibiting activity with target strains other than LAC (Agr subtype I), we measured the production of δ-toxin, for which the gene is embedded in the Agr intracellular effector RNA, RNAIII, in most staphylococci. Production of δ-toxin was measured by RP-HPLC/ESI-MS as described above. Strains LAC (Agr subtype I), A950085 (Agr subtype II), MW2 (Agr subtype III) and A970377 (Agr subtype IV) were used for testing the effect of β-OH-C17-fengycin B on *S. aureus* of different Agr subgroups. Strain 1457 was used for *S. epidermidis*. All strains were diluted 100-fold from a preculture grown in TSB. β-OH-C17-fengycin B dissolved in dimethylsulfoxide (DMSO) was added to each sample to a final concentration of 20 μM and 100 μM. All samples were incubated at 37 °C with shaking for 4 h. Samples were centrifuged and supernatant was collected for RP-HPLC/ESI-MS detection.

Analysis of PVL and α-toxin expression. *S. aureus* strain LAC was diluted 100-fold from a preculture grown in TSB and inoculated into 500 μl TSB. Then, 250 μl of *B. subtilis* culture filtrate was added into the sample. Samples were incubated at 37 °C with shaking for 4 h. Samples were centrifuged in a table-top centrifuge at maximum speed for 5 min; the supernatants were collected and loaded onto 12% SDS–polyacrylamide gel electrophoresis (PAGE) gels, which were run at 160 V for 1 h. Proteins were transferred to nitrocellulose membranes using an iBlot western blotting system. Membranes were incubated with Odyssey blocking buffer for 1 h at room temperature. Anti-α-toxin antibodies (polyclonal rabbit serum; Sigma S7531; dilution 1:5,000) or anti-LukF-PV antibodies (affinity-purified rabbit IgG specific for a peptide region of LukF-PV, produced by GenScript USA and provided by F. DeLeo, NIAID; dilution 1:500) were added to the blocking buffer and membranes were incubated overnight at 4 °C. Then, membranes were washed five times with Tris-buffered saline containing 0.1% Tween-20, pH 7.4, and incubated with Cy5-labelled goat anti-rabbit IgG (diluted 1:10,000 in blocking buffer) in the dark for 1 h at room temperature. Membranes were washed five times with the washing buffer and scanned with a Typhoon TRIO+ variable mode imager.

Preparation of *Bacillus* spores. *B. subtilis* wild-type or isogenic fengycin mutant strains were inoculated from a preculture (1:100) into 1 litre of 2× SG medium[63] and allowed to sporulate for 96 h. Cells were pelleted, washed with water, and resuspended in 20% metrizoic acid (Sigma). Five different concentrations (w/v) of metrizoic acid (60% to 20%) were added stepwise to a 50-ml centrifuge tube to obtain a density gradient. A cell suspension was added to the top of the gradient, and was followed by centrifugation at 40,000*g* for 60 min at 4 °C (as described previously[64]). Spores were found in the middle layers and were collected. They were washed three times with 10 ml water. The total obtained number of viable spores per ml was determined by serial dilution, plating on TSA, and counting of CFU. The total number of heat-resistant spores per ml was determined by submerging the spores in a water bath at 80 °C for 20 min, followed by serial dilution and quantification of CFU per ml as described above.

Mouse intestinal colonization model. In vivo studies were approved by the Institutional Animal Care and Use Committee of the NIAID. Animal work was conducted by certified staff in a facility accredited by the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC). All of the animal work adhered to the institution's guidelines for animal use and followed the guidelines and basic principles in the US Public Health Service Policy on Humane Care and Use of Laboratory Animals, and the Guide for the Care and Use of Laboratory Animals.

All C57BL/6J mice were female and six to eight weeks of age at the time of use. In one set-up, before *S. aureus* was given by oral gavage, mice were pretreated to eradicate the pre-existing intestinal microbiota using an antibiotic mix consisting of ampicillin (1 g l^{-1}), metronidazole (1 g l^{-1}), neomycin trisulfate (1 g l^{-1}) and vancomycin (1 g l^{-1}) in the drinking water. The last day before gavage, antibiotics were omitted from the drinking water. No bacteria could be found in the faeces or intestines of mice for seven days after this treatment in a control experiment. In another set-up, antibiotic pretreatment was omitted. In all set-ups, *S. aureus* strains were grown to mid-exponential growth phase, washed, and resuspended in sterile phosphate-buffered saline (PBS) at 10^8 CFU ml^{-1}. Mice were inoculated by oral gavage with 200 μl of a 10^8 CFU ml^{-1} suspension of the indicated *S. aureus* strains, or 1:1 mixtures of wild-type and isogenic *agr* mutants to reach the same final concentration and volume. For the experiments with strains containing plasmids of the pKX_Δ type, mice received kanamycin (0.2 g l^{-1}) in the drinking water during the experiment to maintain plasmids. For the *B. subtilis* spore competition experiment, oral gavage with 200 μl of spores of wild-type *Bacillus* or its isogenic Δ*fenA* fengycin mutant (10^8 CFU ml^{-1} in sterile PBS) was performed on the day following the *S. aureus* gavage, and repeated every second day thereafter for a total of three times (days 2, 4 and 6). Intestinal colonization was evaluated by quantitative cultures of mouse stool samples and samples from the small and large intestines of mice. In detail, stool was collected and suspended to a final volume of 1 ml of PBS, diluted and plated on TSB agar. Plates were incubated for 24 h at 37 °C, and colonies were enumerated. Moreover, after mice were euthanized seven days after infection, the small and large intestines were collected, resuspended each in 1 ml PBS, and homogenized. Serial dilutions of the homogenates were plated on TSB agar and incubated at 37 °C. Bacterial colonies were enumerated on the following day. In the experiments without antibiotic pretreatment, extracts were plated on MSA plates containing 4 μg ml^{-1} oxacillin (for strain USA300 LAC) or 3 μg ml^{-1} tetracycline (for tetracycline-resistant derivatives of strains ST88 and ST2196), incubated for 48 h at 37 °C, and enumerated.

Statistics. Statistical analysis was performed using GraphPad Prism version 6.05 with one-way or two-way ANOVA, or Fisher's exact test, as appropriate, except for the experiments shown in Figs. 2c, 5b, c, and Extended Data Figs. 2b, c, 5c–f, for which Stata Release 15 and Poisson regression were used, owing to the exclusive presence of 0 values in one group (no variance). For ANOVAs, Tukey post-tests were used, which correct for multiple comparisons using statistical hypothesis testing. All data show the mean and standard deviation (s.d.). All replicates are biological.

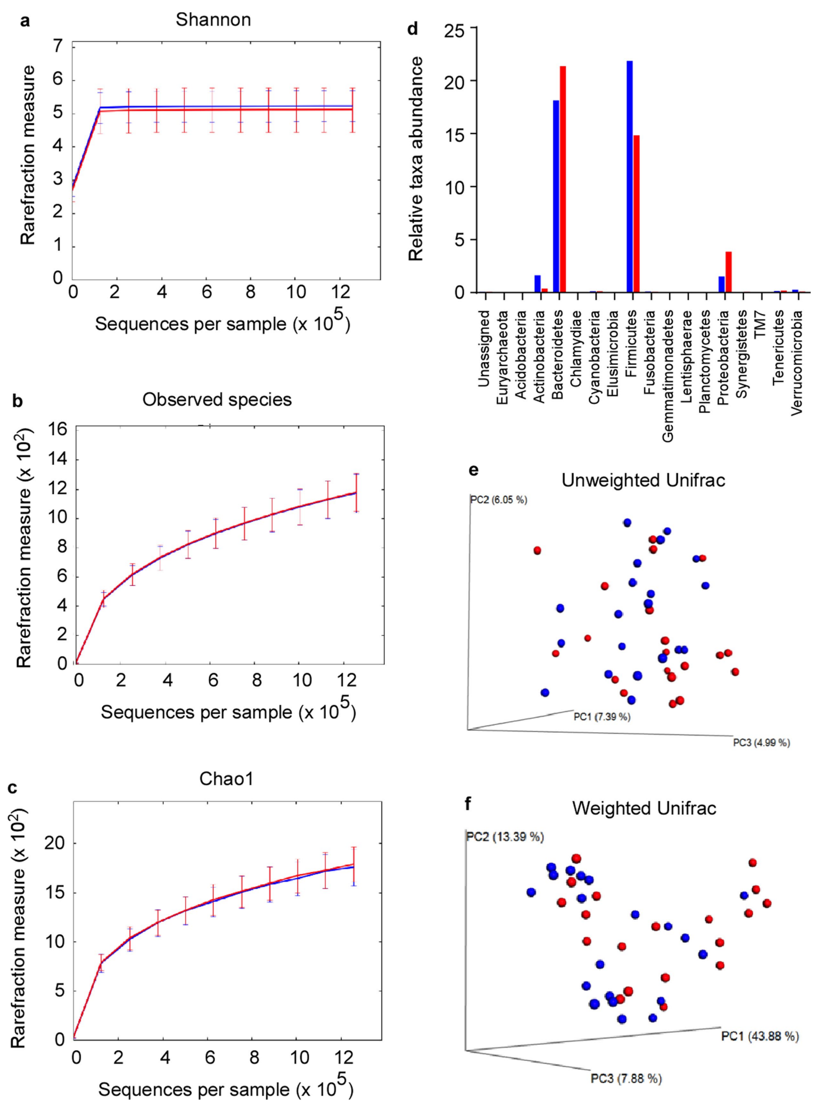
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

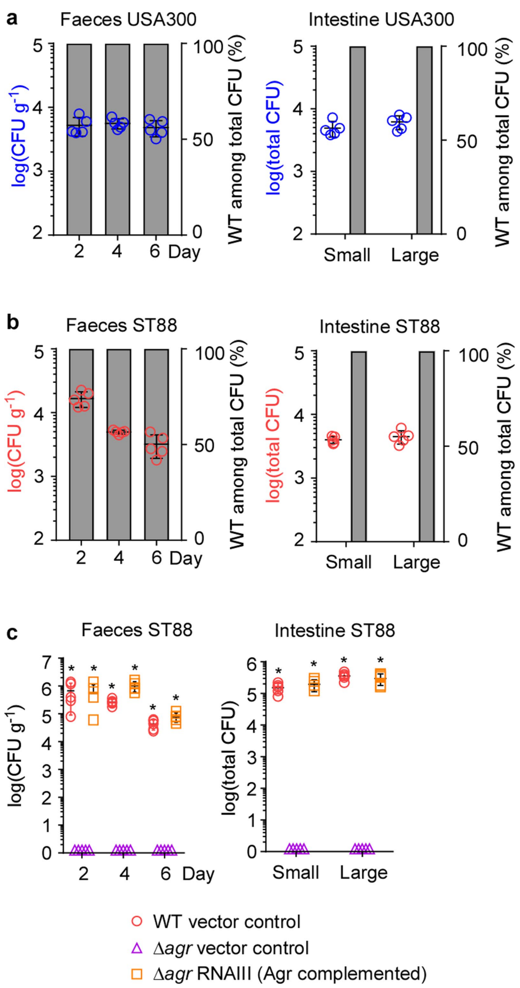
Microbiome sequencing data are available from Bioproject with accession number 483343. All other data generated or analysed during this study are included in the published Article or in the Supplementary Information.

46. Miranda, C. A., Martins, O. B. & Clementino, M. M. Species-level identification of *Bacillus* strains isolates from marine sediments by conventional biochemical, 16S rRNA gene sequencing and inter-tRNA gene sequence lengths analysis. *Antonie van Leeuwenhoek* **93**, 297–304 (2008).
47. Carrel, M., Perencevich, E. N. & David, M. Z. USA300 methicillin-resistant *Staphylococcus aureus*, United States, 2000–2013. *Emerg. Infect. Dis.* **21**, 1973–1980 (2015).
48. Wang, R. et al. Identification of novel cytolytic peptides as key virulence determinants for community-associated MRSA. *Nat. Med.* **13**, 1510–1514 (2007).
49. Gauger, T. et al. Intracellular monitoring of target protein production in *Staphylococcus aureus* by peptide tag-induced reporter fluorescence. *Microb. Biotechnol.* **5**, 129–134 (2012).
50. Queck, S. Y. et al. RNAIII-independent target gene control by the agr quorum-sensing system: insight into the evolution of virulence regulation in *Staphylococcus aureus*. *Mol. Cell* **32**, 150–158 (2008).
51. Monk, I. R., Shah, I. M., Xu, M., Tan, M. W. & Foster, T. J. Transforming the untransformable: application of direct transformation to manipulate genetically *Staphylococcus aureus* and *Staphylococcus epidermidis*. *MBio* **3**, e00277-11 (2012).
52. Luong, T. T. & Lee, C. Y. Improved single-copy integration vectors for *Staphylococcus aureus*. *J. Microbiol. Methods* **70**, 186–190 (2007).
53. Yasbin, R. E. & Young, F. E. Transduction in *Bacillus subtilis* by bacteriophage SPP1. *J. Virol.* **14**, 1343–1348 (1974).

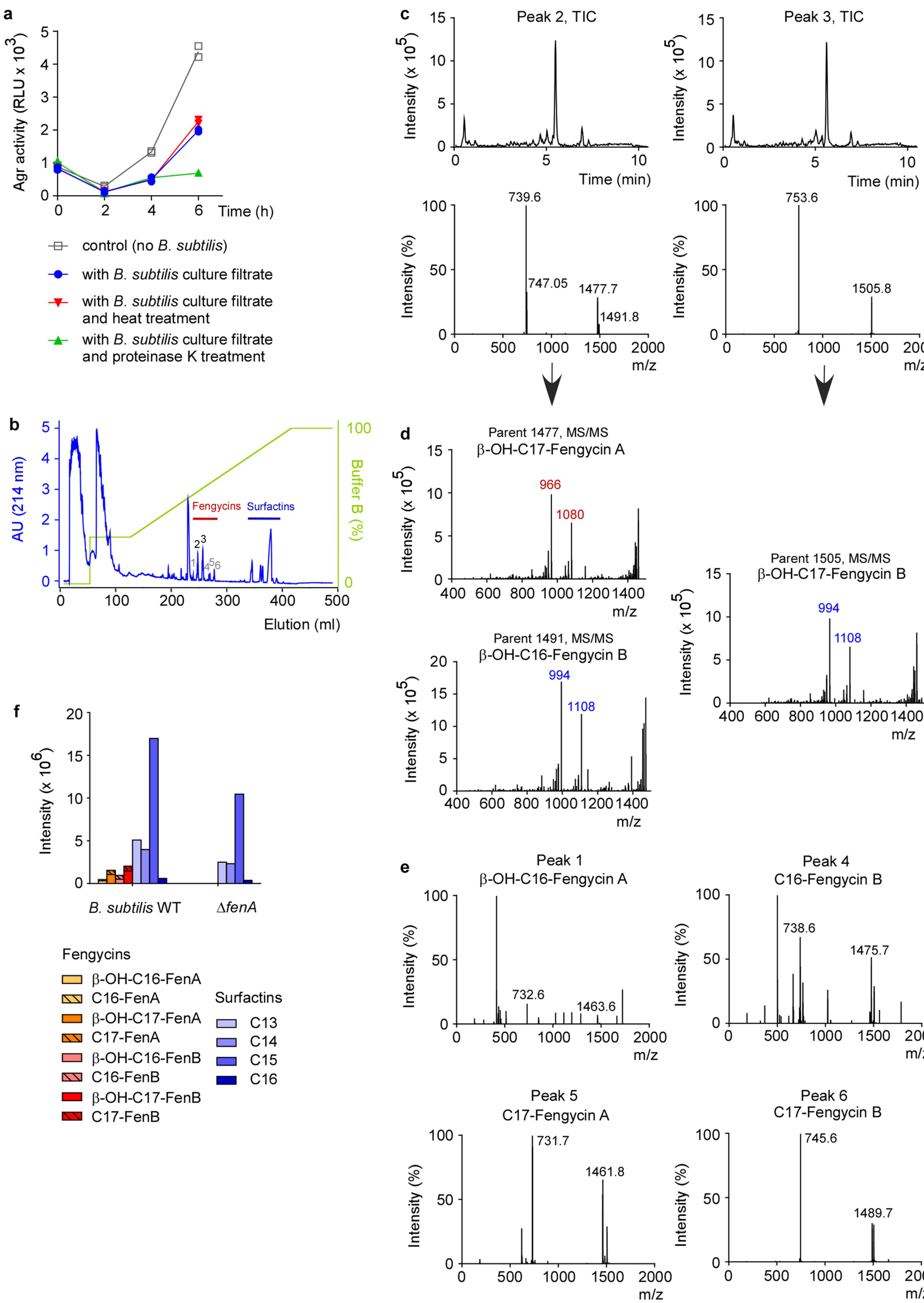
54. Enright, M. C., Day, N. P., Davies, C. E., Peacock, S. J. & Spratt, B. G. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J. Clin. Microbiol*. **38**, 1008–1015 (2000).
55. Francois, P. et al. Rapid *Staphylococcus aureus* agr type determination by a novel multiplex real-time quantitative PCR assay. *J. Clin. Microbiol*. **44**, 1892–1895 (2006).
56. Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J*. **6**, 1621–1624 (2012).
57. Caporaso, J. G. et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
58. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
59. Rideout, J. R. et al. Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2**, e545 (2014).
60. McDonald, D. et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*. **6**, 610–618 (2012).
61. Lozupone, C. A., Hamady, M., Kelley, S. T. & Knight, R. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol*. **73**, 1576–1585 (2007).
62. Joo, H. S. & Otto, M. The isolation and analysis of phenol-soluble modulins of *Staphylococcus epidermidis*. *Methods Mol. Biol*. **1106**, 93–100 (2014).
63. Nicholson, W. L. & Setlow, P. in *Molecular Biological Methods for* Bacillus (eds Harwood, C. R. & Cutting, S. M.) 391–450 (John Wiley, Chichester, 1990).
64. Fukushima, T. et al. Characterization of a polysaccharide deacetylase gene homologue (*pdaB*) on sporulation of *Bacillus subtilis*. *J. Biochem*. **136**, 283–291 (2004).



Extended Data Fig. 1 | Microbiome analysis of *S. aureus* carriers versus non-carriers. The microbiota of $n = 20$ randomly selected *S. aureus* carriers (red) and $n = 20$ non-carriers (blue) were analysed in faecal samples. **a–c**, Rarefaction (species-richness) curves based on 16S rRNA gene sequences. Data are mean ± s.d. **a**, Shannon index. **b**, Observed species. **c**, Chao1 index. **d**, Comparison of relative taxa abundance between *S. aureus* carriers (red) and non-carriers (blue). **e**, **f**, Beta diversity, represented by a principal coordinate analysis plot based on unweighted UniFrac (**e**) and weighted UniFrac (**f**) metrics for samples from *S. aureus* carriers (red) and non-carriers (blue).

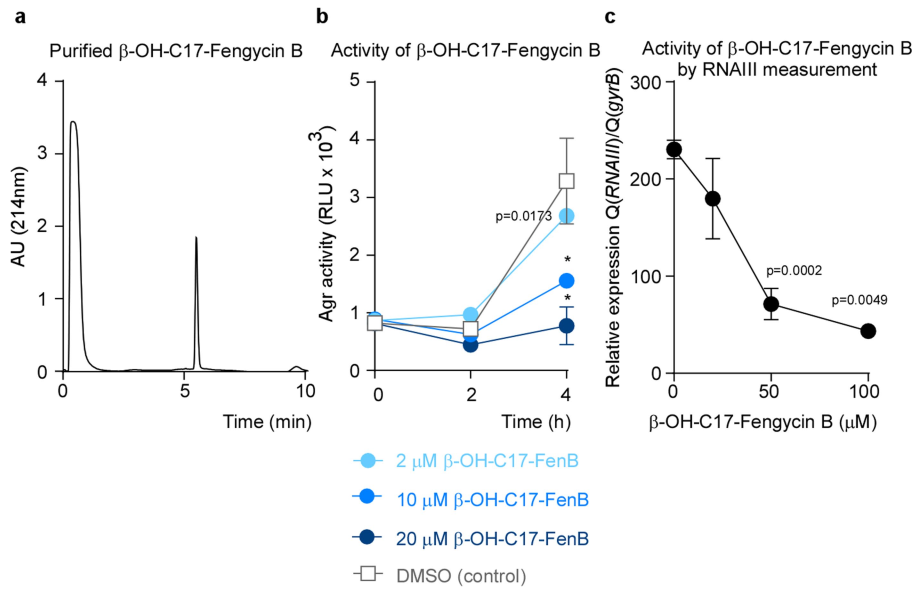


Extended Data Fig. 2 | Quorum-sensing dependence of *S. aureus* intestinal colonization. Data from strains USA300 LAC and ST88 JSNZ. The experimental set-up is the same as in Fig. 2: mice received by oral gavage either 100 μl containing 10^8 CFU ml^{-1} of wild-type *S. aureus* strain USA300 LAC or ST88 JSNZ plus another 100 μl of 10^8 CFU ml^{-1} of the corresponding isogenic *agr* mutant ($n = 5$ per group; competitive experiment shown in **a**, **b**); or 200 μl containing 10^8 CFU ml^{-1} wild-type, isogenic *agr* mutant or Agr (RNAIII)-complemented *agr* mutant ($n = 5$ per group, non-competitive experiment shown in **c**). CFU in the faeces were determined two, four and six days after infection. At the end of the experiment (day seven), CFU in the small and large intestines were determined. **a**, **b**, Competitive experiment. Total obtained CFU are shown as dot plots; also shown are mean ± s.d. Bars show the percentage of wild-type among total determined CFU, of which 100 were analysed for tetracycline resistance that is present only in the *agr* mutant. No *agr* mutants were detected in any experiment; thus, all bars show 100%. Given that 100 isolates were tested, the competitive index wild-type/*agr* mutant in all cases is ≥100. **c**, Non-competitive experiment with genetically complemented strains. Wild-type and isogenic *agr* mutant strains all harboured the pKX$_{\Delta}$16 control plasmid; Agr-complemented strains harboured pKX$_{\Delta}$RNAIII and thus constitutively expressed RNAIII, which is the intracellular effector of Agr. During the experiment, mice received 200 μg ml^{-1} kanamycin in their drinking water to maintain plasmids. Statistical analysis was performed using Poisson regression versus values obtained with the *agr* mutant strains. *$P < 0.0001$. Data are mean ± s.d. Note that no bacteria were found in the faeces or intestines of any mouse receiving *S. aureus* Δ*agr* with vector control. The corresponding zero values are plotted on the *x* axis of the logarithmic scale.

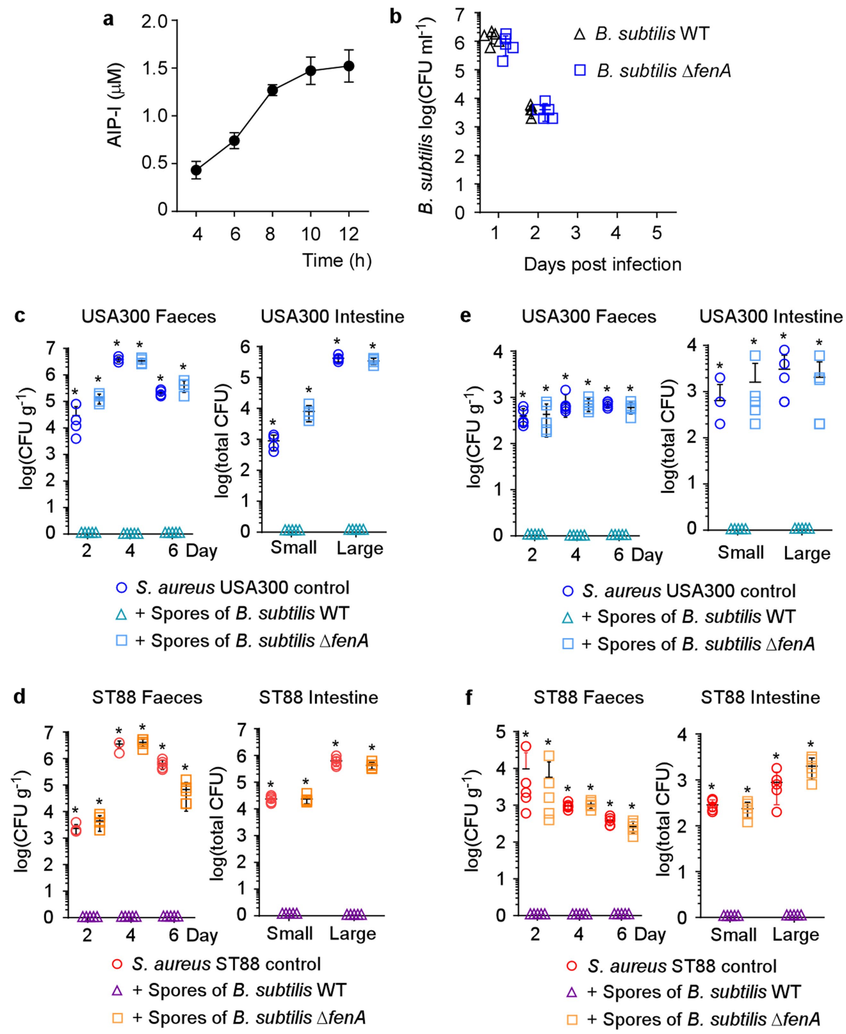


Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Analysis of Agr-inhibitory substances. **a**, Influence of heat and proteases on Agr inhibition. *B. subtilis* culture filtrate was subjected to heat (95 °C for 20 min) or digestion with proteinase K (50 μg ml^{-1}, 37 °C, 1 h) and the effect on inhibition of Agr activity was measured using the luminescence assay with the USA300 P3–*luxABCDE* reporter strain (see Fig. 3a). RLU, relative light units. The experiment was performed with $n = 2$ independent biological samples. Lines connect the means. (The observed additional suppression of Agr activity in the proteinase-K-treated sample at 6 h, compared with the *B. subtilis* culture filtrate sample, is expected owing to proteolytic inactivation of intrinsic AIP.) **b**, Preparative RP chromatography of *B. subtilis* culture filtrate to determine the Agr-inhibiting substance. The peaks labelled 2 and 3 showed substantial Agr-inhibiting activities in the Agr-activity assay and were identified as fengycins using subsequent RP-HPLC/ESI-MS and MS/MS analysis (see **c**, **d**). The peaks labelled 1 and 4–6 also contained fengycin species (see **e**). AU, arbitrary units. The applied gradient (% buffer B) is shown in green. **c**, Fractions corresponding to Agr-inhibitory peaks 2 and 3 from the preparative RP run (**b**) were subjected to RP-HPLC/ESI-MS. Top, total ion chromatograms (TICs) of the RP-HPLC/ESI-MS runs; bottom, ESI mass spectrogram of the major peaks. **d**, MS/MS analysis of the peak 2 and 3 fractions. Peaks that are characteristic of a given fengycin subtype (A or B in this case) are marked in colour. 'Parent' refers to the relevant numbered peak in the spectrograms above. **e**, Analysis of further fengycin-containing fractions. Peaks 1, 4, 5 and 6 from the preparative RP run (**b**) were also found to contain fengycin species as determined by subsequent RP-HPLC/ESI-MS analysis. Shown are the mass spectrograms of the major peaks of those runs and the tentative characterization for fengycin type. The preparative and analytical chromatography and RP-HPLC/ESI-MS analyses (as shown in **b**, **d**) were repeated multiple (more than ten) times for fengycin purification, with similar results. MS/MS analyses were not repeated. **f**, Analysis of fengycin and surfactin lipopeptide expression by the *B. subtilis* wild-type strain and its isogenic Δ*fenA* mutant.



Extended Data Fig. 4 | Assessment of purity and functionality of purified β-OH-C17-fengycin B. a, RP-HPLC run. **b**, Agr inhibition at different concentrations in the luminescence assay. RLU, relative light units. Statistical analysis was by two-way ANOVA with Tukey's post-test. Comparisons shown are those versus DMSO control. **c**, Agr inhibition as measured by inhibition of expression of RNAIII by qRT–PCR. *$P < 0.0001$ (one-way ANOVA with Tukey's post-test; comparisons shown are those versus 0 μM value). The experiments in **b**, **c** were performed with $n = 3$ independent biological samples. Data are mean ± s.d.



Extended Data Fig. 5 | Inhibition of *S. aureus* colonization by dietary fengycin-producing *Bacillus* spores in a mouse model. a, Concentration of AIP-I during *S. aureus* growth. Strain LAC (USA300) was grown in TSB, and AIP-I concentrations were measured by RP-HPLC/ESI-MS. Calibration was performed using synthetic AIP-I. The detection limit of this assay is around 0.3 μM. The experiment was performed with $n = 3$ independent biological samples. Data are mean ± s.d. **b**, *B. subtilis* colonization kinetics in the mouse intestinal colonization experiment. Mice ($n = 5$) received 200 μl of a 10^8 CFU ml^{-1} suspension of wild-type *B. subtilis* or Δ*fenA* mutant spores by oral gavage; CFU in the faeces were analysed up to five days afterwards. Data are mean ± s.d. **c–f**, Inhibition mouse model with strains USA300 LAC and ST88 JSNZ. The experimental set-up was as shown in Fig. 5a. In brief, $n = 4$ or 5 mice per group received 200 μl of 10^8 CFU ml^{-1} *S. aureus* strains USA300 LAC or ST88 JSNZ by oral gavage. On the next day and every following second day, the mice received 200 μl of 10^8 CFU ml^{-1} spores of wild-type *B. subtilis* or its isogenic *fenA* mutant, also by oral gavage. CFU in the faeces were determined two, four and six days after infection. At the end of the experiment (day seven), CFU in the small and large intestines were determined. The experiment was performed with (**c**, **d**) or without (**e**, **f**) antibiotic pretreatment. Statistical analysis was performed using Poisson regression versus values obtained with wild-type *B. subtilis* spore samples. *$P < 0.0001$. Data are mean ± s.d. Note that no *S. aureus* were found in the faeces or intestines of any mouse challenged with any *S. aureus* strain that also received *Bacillus* wild-type spores. The corresponding zero values are plotted on the *x* axis of the logarithmic scale.

Extended Data Table 1 | Fengycin production and Agr-inhibition potency of *Bacillus* faecal isolates

Isolate Code	*Bacillus* species*	β - OH-C16-FenA	C16-FenA	β - OH-C17-FenB	C17-FenB	β - OH-C17-FenA	C17-FenA	β - OH-C16-FenB	C16-FenB	% Agr inhibition†	Total Fengycin Concentration
10	*licheniformis*	100	48	65	80	33	33	51	32	98	442
14	*subtilis*	82	27	104	138	52	111	186	0	97	700
15	*amyloliquefaciens*	7	212	212	83	152	56	274	81	92	1076
16	*sonorensis*	5,833	437	984	288	2,751	1,248	1,760	1,090	98	14390
18	*subtilis*	107	0	58	118	0	0	53	23	97	359
19	*licheniformis*	0	61	0	75	13	0	149	31	97	329
21	*sonorensis*	0	85	159	48	0	104	18	33	95	447
26	*megaterium*	47	0	112	8	33	134	48	23	98	405
30	*subtilis*	0	0	63	28	109	77	87	43	96	407
31	*sonorensis*	23	0	52	0	157	15	0	0	97	246
32	*sonorensis*	23	118	145	0	0	0	74	59	94	418
33	*licheniformis*	89	119	70	99	98	102	150	0	96	727
35	*licheniformis*	0	0	98	0	149	72	28	132	96	479
36	*pumilus*	0	0	10	0	0	136	144	0	98	290
37	*subtilis*	116	232	67	0	15	45	0	114	96	589
38	*subtilis*	59	167	215	0	79	114	50	212	93	896
39	*licheniformis*	150	117	249	0	241	271	226	230	96	1484
40	*subtilis*	753	174	1,260	1,124	1,548	538	841	821	80	7058
41	*subtilis*	2,298	860	1,777	0	4,524	816	1,563	1,957	91	13796
42	*sonorensis*	34	43	0	0	41	57	88	0	96	263
43	*sonorensis*	477	0	2,488	816	1,667	604	1,216	858	96	8126
45	*sonorensis*	98	0	24	25	0	0	0	39	99	187
47	*pumilus*	342	14	237	0	0	108	105	0	97	806
48	*subtilis*	0	0	49	0	7	0	0	0	94	56
49	*subtilis*	5,007	828	979	0	3,147	0	2,027	1,210	97	13200
50	*subtilis*	429	0	933	0	1,065	354	614	526	90	3922
51	*subtilis*	712	185	911	208	1,187	665	863	586	88	5316
52	*subtilis*	9,630	1,571	923	0	2,365	1,775	4,690	1,099	95	22052
53	*licheniformis*	0	0	0	0	0	0	58	25	95	83
55	*subtilis*	45	43	113	12	94	0	0	216	96	523
56	*sonorensis*	167	53	0	0	83	201	109	58	98	671
57	*subtilis*	127	104	166	301	61	96	98	246	96	1200
58	*sonorensis*	498	510	841	0	967	222	0	0	94	3039
59	*subtilis*	0	0	1,008	0	0	297	715	488	82	2508
61	*subtilis*	124	84	21	0	160	39	42	335	97	805
62	*licheniformis*	93	128	7	48	40	188	0	153	97	657
63	*amyloliquefaciens*	4,153	380	975	0	2,584	1,047	1,975	1,084	95	12197
64	*sonorensis*	133	91	187	230	29	387	62	34	95	1154
65	*subtilis*	9	126	0	0	35	90	239	36	97	535
66	*sonorensis*	254	0	0	0	377	0	131	0	97	762
67	*subtilis*	215	84	200	0	0	0	155	168	97	821
68	*subtilis*	41	12	144	0	0	0	115	182	93	494
69	*subtilis*	266	27	236	290	132	0	54	195	98	1200
70	*subtilis*	54	55	390	0	196	0	187	112	94	994
71	*subtilis*	14,881	4,578	88,106	34,879	39,939	16,967	42,502	23,859	97	265710
74	*pumilus*	157	56	54	14	0	62	177	41	95	560
75	*licheniformis*	281	22	40	0	0	292	125	204	95	964
76	*subtilis*	124	0	74	0	0	101	82	192	97	573
77	*sonorensis*	0	0	0	25	0	0	93	13	97	131
78	*amyloliquefaciens*	73	99	0	0	79	17	176	13	94	458
79	*amyloliquefaciens*	10	0	0	0	0	180	51	63	97	304
80	*subtilis*	1,741	322	4,222	1,105	3,327	910	2,207	1,529	91	15363
81	*subtilis*	1,739	426	5,073	1,579	3,371	998	3,241	1,933	87	18361
82	*subtilis*	1,002	0	3,413	921	1,998	0	1,710	992	91	10037
83	*licheniformis*	356	0	536	4	83	201	107	52	97	1338
85	*subtilis*	52	0	49	0	84	161	0	132	95	479
87	*subtilis*	1,327	312	3,931	0	2,763	667	2,624	1,167	98	12790
88	*pumilus*	101	0	215	156	313	367	276	216	96	1643
89	*subtilis*	105	59	266	0	0	0	51	38	93	519
91	*subtilis*	325	91	493	17	120	290	302	186	98	1825
92	*subtilis*	254	0	234	156	275	0	154	66	96	1140
93	*subtilis*	493	273	91	342	204	287	277	435	96	2402
94	*subtilis*	876	115	134	37	384	190	445	529	91	2712
95	*amyloliquefaciens*	351	175	157	146	110	225	196	197	97	1557
97	*subtilis*	1,845	912	4,714	1,310	3,686	1,492	2,484	1,557	86	18000
98	*subtilis*	1,367	804	3,572	1,512	2,803	1,511	2,005	1,626	93	15200
99	*subtilis*	77	375	705	0	28	170	115	77	91	1547
100	*amyloliquefaciens*	81	117	337	267	105	166	237	502	81	1811
103	*pumilus*	249	45	350	207	162	249	536	279	98	2077
104	*subtilis*	293	77	105	269	75	509	286	17	99	1632
106	*subtilis*	978	796	4,415	3,419	1,874	1,378	2,478	1,527	97	16866
107	*subtilis*	423	199	322	150	520	517	384	160	98	2675
108	*pumilus*	224	114	397	0	229	211	413	43	99	1631
110	*subtilis*	140	77	317	286	404	127	117	139	95	1607
111	*pumilus*	184	104	319	67	96	212	294	120	93	1395
112	*subtilis*	470	183	1,412	637	1,211	732	950	655	96	6251
113	*pumilus*	463	202	276	204	211	0	156	137	95	1650
115	*subtilis*	268	232	297	62	313	410	713	88	97	2382
116	*subtilis*	352	205	369	0	172	298	561	350	96	2306
117	*subtilis*	143	104	716	328	0	149	97	266	98	1803
118	*subtilis*	163	34	1,788	0	0	155	306	42	93	2488
119	*amyloliquefaciens*	604	256	947	258	0	350	361	104	93	2880
121	*subtilis*	503	151	0	364	63	84	174	162	98	1502
122	*subtilis*	152	311	24	165	86	213	296	83	91	1329
123	*subtilis*	8,801	3,540	23,045	10,778	16,465	6,428	16,077	10,706	98	95839
124	*subtilis*	106	139	316	168	64	175	195	103	86	1266
125	*subtilis*	0	0	0	158	0	318	290	0	95	765
126	*subtilis*	288	157	211	110	428	421	185	112	87	1913
127	*subtilis*	478	193	103	435	240	303	551	132	96	2434
128	*subtilis*	177	156	228	118	96	276	426	48	97	1525
129	*pumilus*	249	37	0	162	0	224	0	144	97	816
130	*subtilis*	1,267	0	4,668	1,266	2,940	1,168	2,158	1,863	88	15332
131	*subtilis*	4,164	496	513	307	1,226	671	1,948	662	83	9986
132	*subtilis*	441	291	491	391	200	562	381	279	95	3036
134	*subtilis*	6,647	750	1,011	0	3,143	1,166	2,829	953	97	16498
136	*subtilis*	751	233	2,569	773	1,813	890	1,701	1,010	89	9740
137	*subtilis*	1,572	328	3,297	1,036	2,221	627	2,266	0	82	11347
138	*pumilus*	288	311	232	11	415	236	236	403	85	2133
139	*subtilis*	1,898	709	7,830	2,453	5,328	2,073	3,880	2,210	98	26381
140	*sonorensis*	0	106	217	815	225	0	103	3,987	91	5454
141	*thuringiensis*	258	26	325	230	0	264	0	0	87	1102
142	*pumilus*	422	159	262	0	402	335	639	251	87	2471
143	*subtilis*	250	37	210	16	64	68	293	89	83	1027
144	*sonorensis*	110	208	134	191	0	304	361	91	87	1399
	B. subtilis ZK3814	2,563	1,781	17,078	5,725	11,963	4,448	5,444	4,902	94	53904

The table shows intensity values from the integration of *m/z* peaks associated with the specific fengycin species, as obtained by RP-HPLC/ESI-MS. The two most abundant peaks, corresponding to double- and triple-charged ions, were used for the integration. Values are in nM, obtained by calibration using weighed and diluted aliquots of the *Bacillus* lipopeptide surfactin.

**Bacillus* species were determined by sequencing 16S RNA encoding DNA, as specified in the Methods.

†The percentage of Agr inhibition was determined by dividing the 4-h value obtained in the luminescence assay for the sample (using 100 μl of culture filtrate) by that obtained for the control, and multiplying by 100.

Extended Data Table 2 | Analysis of previous microbiome studies for correlation between the presence of *S. aureus* and *B. subtilis* in the human intestinal tract

Study ID	Study Name	Samples	Only *B. subtilis*	Only *S. aureus*	Both	Neither
ERP012803	American Gut Project	6635	1 (0.015%)	304 (4.58%)	0	6330 (95.4%)
ERP011001	Human gut bacteria that rescue growth and metabolic defects transmitted by microbiota from undernourished children	1732	408 (23.61%)	70 (4.05%)	71 (4.11%)	1179 (68.23%)
ERP005437	16S sequencing of Malawian children	1515	118 (7.79%)	6 (0.4%)	4 (0.26%)	1387 (91.55%)
SRP049113	Human gut microbiota from the ALADDIN study	664	2 (0.30%)	61 (9.19%)	7 (1.05%)	594 (89.46%)
ERP019564	Role of Gut Microbiota in Pathophysiology of Parkinson's Disease	481	8 (1.66%)	7 (1.45%)	0	466 (96.88%)
SRP073172	DNA from FIT can replace stool for microbiota-based colorectal	408	63 (15.44%)	71 (17.40%)	99 (24.26%)	175 (42.89%)
SRP068240	Human feces metagenome 16s rDNA sequencing	350	52 (14.85%)	189 (54%)	89 (25.43%)	20 (5.71%)
SRP064846	Homo sapiens fecal microbiome transplant	271	20 (7.38%)	47 (17.34%)	6 (2.21%)	198 (73.06%)
SRP065497	Human gut environment Targeted loci environmental	270	54 (20%)	8 (2.96%)	19 (7.04%)	189 (70%)
ERP021093	Gut microbiome from patients obtained by 16s rRNA sequencing.	268	88 (32.84%)	14 (5.22%)	57 (21.27%)	109 (40.67%)
ERP010229	Gut microbial succession follows acute secretory diarrhea in humans	260	12 (4.62%)	92 (35.38%)	122 (46.92%)	34 (13.08%)
ERP010458	Gut microbiota of stroke patients differentiates from healthy controls	233	3 (1.29%)	32 (13.73%)	4 (1.72%)	194 (83.26%)

We included in our analysis all studies found on the EBI Metagenomics website (https://www.ebi.ac.uk/metagenomics/) that had more than 200 participants (independent samples) and which used Illumina Miseq instruments. We pooled raw 16S rRNA sequencing data from the EBI Metagenomics website, and used taxonomic assignment (TSV) files for analysis. The number of sequence reads was used to analyse how many samples contained *S. aureus* or *B. subtilis*. Samples with a read number of more than 0 were defined as colonized. When there were no reads, samples were designated as noncolonized.

ARTICLE

https://doi.org/10.1038/s41586-018-0621-1

Transcription factor dimerization activates the p300 acetyltransferase

Esther Ortega[1], Srinivasan Rengachari[1,5], Ziad Ibrahim[1,2], Naghmeh Hoghoughi[3], Jonathan Gaucher[1,6], Alex S. Holehouse[4], Saadi Khochbin[3] & Daniel Panne[1,2]*

The transcriptional co-activator p300 is a histone acetyltransferase (HAT) that is typically recruited to transcriptional enhancers and regulates gene expression by acetylating chromatin. Here we show that the activation of p300 directly depends on the activation and oligomerization status of transcription factor ligands. Using two model transcription factors, IRF3 and STAT1, we demonstrate that transcription factor dimerization enables the *trans*-autoacetylation of p300 in a highly conserved and intrinsically disordered autoinhibitory lysine-rich loop, resulting in p300 activation. We describe a crystal structure of p300 in which the autoinhibitory loop invades the active site of a neighbouring HAT domain, revealing a snapshot of a *trans*-autoacetylation reaction intermediate. Substrate access to the active site involves the rearrangement of an autoinhibitory RING domain. Our data explain how cellular signalling and the activation and dimerization of transcription factors control the activation of p300, and therefore explain why gene transcription is associated with chromatin acetylation.

Signals that emanate from cellular receptors ultimately lead to changes in gene expression that drive cellular change and organismal development. Gene expression is typically controlled through the coordinated activity of DNA-binding transcription factors, chromatin regulators and the general transcription machinery. For instance, in the innate immune system, pattern recognition receptors recognize and engage with various pathogen-associated molecular patterns[1], and subsequently bind to adaptor proteins such as STING (stimulator of interferon genes). These adaptor proteins engage the latent DNA-binding transcription factor interferon (IFN) regulatory factor 3 (IRF3) and enable recruitment and activation of the non-canonical IκB kinase TBK1[1]. TBK1 then phosphorylates IRF3 in a C-terminal motif, resulting in the removal of autoinhibition, dimerization and adaptor displacement[2,3]. Activated IRF3 dimers bind to p300/CBP (where CBP is CREB-binding protein; p300 and CBP are also known as KAT3B and KAT3A, respectively) to stimulate chromatin acetylation and gene expression of the antiviral type I IFNs IFNα and IFNβ[3–5]. Type I IFNs are secreted and bind to specific cell-surface IFN receptors, which results in the activation of Janus kinase–signal transducers and activators of transcription (JAK–STAT) signalling[6]. The activated, tyrosine-phosphorylated STATs then dimerize, translocate to the nucleus and bind to p300/CBP to stimulate the transcription of IFN-stimulated genes[7].

p300/CBP are known to interact with more than 400 binding partners including the basal transcription machinery[8]. The large protein interactome of p300/CBP results in near-universal recruitment of these HATs to enhancers, and p300 occupancy has been used to identify enhancers genome-wide[9,10]. p300/CBP catalyses the acetylation of histone H3K27 to form H3K27ac, a modification that is considered an 'activation' mark[11]. However, recruitment of p300/CBP does not always correlate with gene activation and is occasionally associated with repression[12–16]. A large number of chromatin regions that bind p300/CBP therefore do not contain this canonical H3K27ac modification, which indicates that HAT activity at such sites is blocked[15,17]. Therefore, it is a major challenge to understand the mechanism that enables switching between inactive and active states of p300/CBP on enhancers, and to causally link cellular signalling to the recruitment of p300/CBP, the regulation of HAT activity and the establishment of repressed, poised and active chromatin.

Here we have investigated how the activation and oligomerization status of p300 transcription factor ligands such as IRF3 and STAT1 affects the catalytic activity of p300. We found that the kinase-activated and dimeric, but not the inactive or monomeric, variants of these transcription factors support robust p300 HAT activation. We demonstrate that transcription factor dimerization enables p300 *trans*-autoacetylation in a lysine-rich, intrinsically disordered autoinhibitory loop (AIL) in the HAT domain that serves as a 'pseudosubstrate' and is important for regulating the HAT activity of p300[18]. A crystal structure of the core domain of p300 provides a snapshot of a potential *trans*-autoacetylation reaction intermediate in which the AIL projects into the active site of a neighbouring p300 molecule. As HAT activation is closely linked to transcription factor activation, these results causally relate cellular signalling to the activation and DNA targeting of a chromatin modifier and provide mechanistic insights into the long-standing and general correlation between an active, acetylated chromatin structure and gene transcription.

Transcription factor dimerization activates p300

To explore whether p300 is activated by signal-dependent IRF3 dimerization, we produced three recombinant IRF3 species: inactive monomers (IRF3); active, TBK1-phosphorylated IRF3 dimers (pIRF3); and a truncation mutant that lacked the C-terminal autoinhibitory element (IRF3ΔC) (Extended Data Fig. 1a, b). Truncation of the C-terminal autoinhibitory element allows for p300/CBP binding but abolishes IRF3 dimerization[19]. We confirmed the oligomerization status by gel filtration chromatography (Extended Data Fig. 1b), and investigated the effect of IRF3 activation and oligomerization status on the autoacetylation of p300s in the presence of [^{14}C]acetyl coenzyme A

[1]European Molecular Biology Laboratory, Grenoble, France. [2]Leicester Institute of Structural and Chemical Biology, Department of Molecular and Cell Biology, University of Leicester, Leicester, UK. [3]CNRS UMR 5309, INSERM U1209, Université Grenoble Alpes, Institute for Advanced Biosciences, Grenoble, France. [4]Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, St. Louis, MO, USA. [5]Present address: Max Planck Institute for Biophysical Chemistry, Department of Molecular Biology, Göttingen, Germany. [6]Present address: Université Grenoble Alpes, INSERM U1042, HP2 Laboratory, Grenoble, France. *e-mail: daniel.panne@le.ac.uk

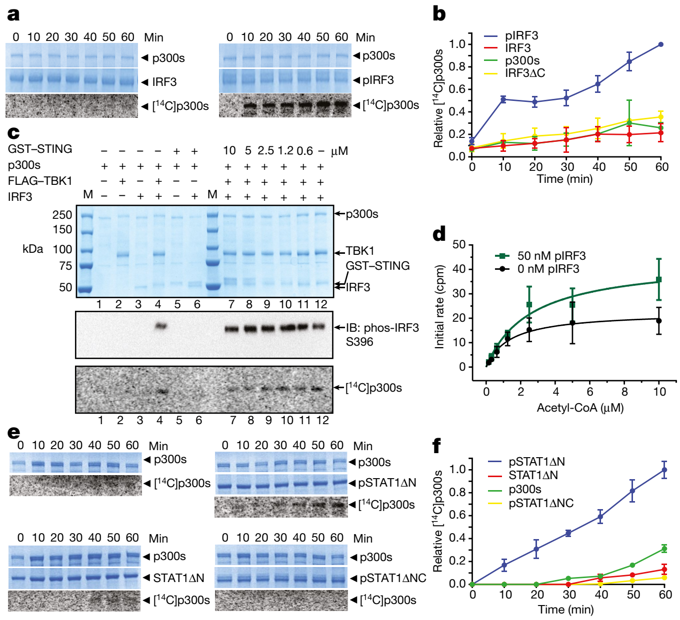


Fig. 1 | Transcription factor dimerization activates p300. a, p300s was incubated for the indicated times in the presence or absence of inactive, monomeric IRF3 or TBK1-phosphorylated, dimeric pIRF3. Samples were analysed by SDS–PAGE followed by Coomassie staining and autoradiography. Representative data of three independent experiments are shown. **b**, Quantification of the autoacetylation of p300s. **c**, p300 is activated by TBK1-mediated IRF3 phosphorylation. p300s was incubated with recombinant GST–STING, TBK1 and IRF3 in the presence of ATP and [^{14}C]acetyl-CoA. Top, Coomassie-stained SDS–PAGE gel. Middle, analysis of IRF3 phosphorylation on S396 using immunoblotting. Bottom, autoradiography. Representative data of three independent experiments are shown. **d**, HAT scintillation proximity assay. The degree of histone H4 substrate acetylation was quantified using scintillation counting. **e**, As in **a** but using inactive, monomeric STAT1ΔN or activated, dimeric pSTAT1ΔN. Activated, dimeric pSTAT1ΔNC that lacks the C-terminal TAD did not stimulate the autoacetylation of p300s. Samples were analysed as in **a**. Representative data of three independent experiments are shown. **f**, Quantification of the autoacetylation of p300s. Intensity values were normalized by dividing by the maximum autoacetylation signal obtained after 60 min. Error bars shown in **b**, **d** and **f**: three independent experiments were performed, data are mean ± s.d. Data analysis and plotting was performed with GraphPad Prism 7.0. For gel source data, see Supplementary Fig. 1.

(acetyl-CoA). p300s is a short p300 construct that spans from the TAZ1 to the NCBD/IBiD domains and contains a deletion of the flexible N- and C-terminal regions (Extended Data Fig. 4a). p300s autoacetylated slowly in the absence of IRF3 (Extended Data Fig. 1c). The inclusion of inactive, monomeric IRF3 or IRF3ΔC did not modify HAT activity (Fig. 1a, Extended Data Fig. 1c). By contrast, the inclusion of active, TBK1-phosphorylated IRF3 dimers (pIRF3) resulted in a rapid burst of autoacetylation followed by a gradual increase of acetylated p300s (Fig. 1a, b). As IRF3ΔC did not support p300 HAT activation, IRF3 dimerization and not solely p300 binding is essential for HAT activation.

p300 HAT activation was directly dependent on the TBK1-mediated phosphorylation of IRF3 on Ser396, a critical residue for IRF3 activation and dimerization[2,3]. Only when both TBK1 and IRF3 were included in the reaction did we observe p300 activation (Fig. 1c, lane 4). We observed only a modest stimulatory effect of the adaptor protein STING (Fig. 1c, lanes 7–12), probably owing to the relatively large amounts of TBK1, which is already active and phosphorylates IRF3 even in the absence of STING[20]. We conclude that IRF3 phosphorylation by TBK1 and its dimerization are required for p300 HAT activation.

To analyse the effect of pIRF3 on p300 activation and acetylation of the histone substrate, we established a scintillation proximity HAT assay. We incubated saturating amounts of a biotinylated histone H4 substrate peptide with p300s in the absence or the presence of equimolar pIRF3 and increasing concentrations of [^{3}H]acetyl-CoA (Fig. 1d). pIRF3 stimulated p300 histone-substrate acetylation, as determined by the increased rate of H4 acetylation obtained in the presence of pIRF3 ($V_{max} = 43.8 \pm 5.3$ cpm min^{-1} as compared to $V_{max} = 22.5 \pm 2.8$ cpm min^{-1} in the absence of pIRF3). These data indicate that pIRF3 not only stimulates p300 autoacetylation and activation, but also stimulates more efficient acetylation of the histone substrate.

We also investigated the effect of STAT1 on p300 activation. STATs are activated in response to cytokine receptor engagement and Janus kinase activation[21]. JAK-mediated phosphorylation of STAT1 on Tyr701 induces dimerization and translocation to the nucleus, where STAT1 binds to DNA elements to regulate gene expression. STAT1 contains a C-terminal transactivation domain (TAD) through which it interacts with p300/CBP[7]. A naturally occurring splice variant, STAT1β, lacks the TAD and acts in a dominant negative manner[22]. Structures of the active, STAT1 Tyr701-phosphorylated dimer bound to DNA, as well as the STAT1 TAD bound to the TAZ2 domain of CBP, have been determined previously[23,24].

To understand the effect of STAT1 activation and oligomerization status on p300 activity, we produced STAT1ΔN that lacked the N-domain and STAT1ΔNC lacking the N-domain and the TAD as non-phosphorylated monomers or as Tyr701-phosphorylated dimers (Extended Data Fig. 1e–h). We found that p300s autoacetylated slowly in the absence of STAT1, and that the addition of non-phosphorylated, monomeric STAT1ΔN did not stimulate p300s autoacetylation beyond background levels (Fig. 1e, f). By contrast, the addition of Tyr701-phosphorylated STAT1ΔN (pSTAT1ΔN) dimers to p300s resulted in a rapid increase of p300 autoacetylation. Activation required the C-terminal TAD of STAT1, as the addition of a Tyr701-phosphorylated STAT1 dimer (pSTAT1ΔNC) that lacked the TAD did not stimulate p300 autoacetylation (Fig. 1e, f).

STAT1 dimerization, and not solely interaction with the TAZ2 domain, is required for the activation of p300. Unphosphorylated, monomeric STAT1ΔN, which contains the TAD and is able to interact with the TAZ2 domain of CBP, did not stimulate p300 activity. However, stimulation with STAT1 was not as potent compared with that of IRF3, possibly because our STAT1 preparation is unphosphorylated on Ser727, which is required for maximal gene activation[25]. Together, our data are consistent with a model in which the AIL peptide serves as an intramolecular 'pseudosubstrate' and a competitive HAT inhibitor[18]. Dimeric ligands such as pIRF3 and pSTAT1 allow p300 activation by bringing two molecules together to enable *trans*-autoacetylation of the AIL, which in turn relieves autoinhibition and enables more efficient entry of substrates into the HAT active site.

Structure of p300 adopts an AIL-swap conformation

To further understand the role of the AIL in the regulation of these structural transitions, we crystallized the hypoacetylated form of the catalytic core of p300 comprising the bromo-RING-PHD-HAT domains (BRP-HAT) that contained the AIL. Crystals were obtained using a similar protocol as published previously[26]. Crystals diffracted to a minimal Bragg spacing of 3.1 Å and we determined the structure by molecular replacement (Extended Data Table 1). The crystal form contained four p300 molecules in the asymmetric unit (Extended Data Fig. 2). Comparison with our previous structure[26] showed that the bromo-PHD-HAT domains overlay well on each other with a root-mean-square deviation (r.m.s.d.) of approximately 1 Å. However, the RING domains were not visible in the initial electron density map. Anomalous difference density maps showed a density peak for the zinc atom of the RING domain, but it was not at the expected location. Manual repositioning enabled the correct placement of the RING domains into the new position and the refinement of the structure (Fig. 2a, Extended Data Fig. 3).

The p300 molecules show an antiparallel arrangement of the BRP-HAT domains (Extended Data Fig. 2a). In this configuration, the HAT

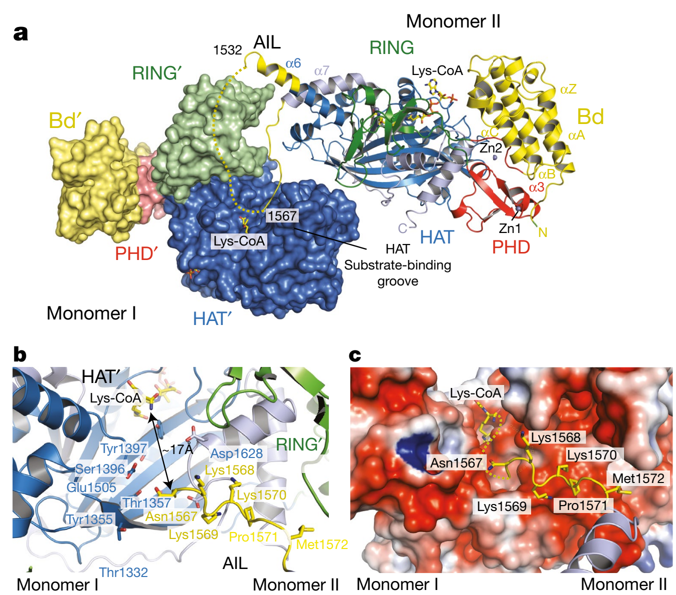


Fig. 2 | The structure of p300 adopts a AIL-swap conformation. **a**, Monomer I is surface-rendered and monomer II is shown as a cartoon. The AIL loop from monomer II is shown in yellow. The AIL lies near the HAT substrate-binding groove of monomer I. A disordered segment of the AIL is shown as a dotted line. **b**, Close-up view of the residues of the AIL loop from monomer II and residues of monomer I in the substrate-binding pocket. **c**, Binding of the positively charged AIL is mediated by interactions with negatively charged residues in the substrate-binding pocket of the HAT.

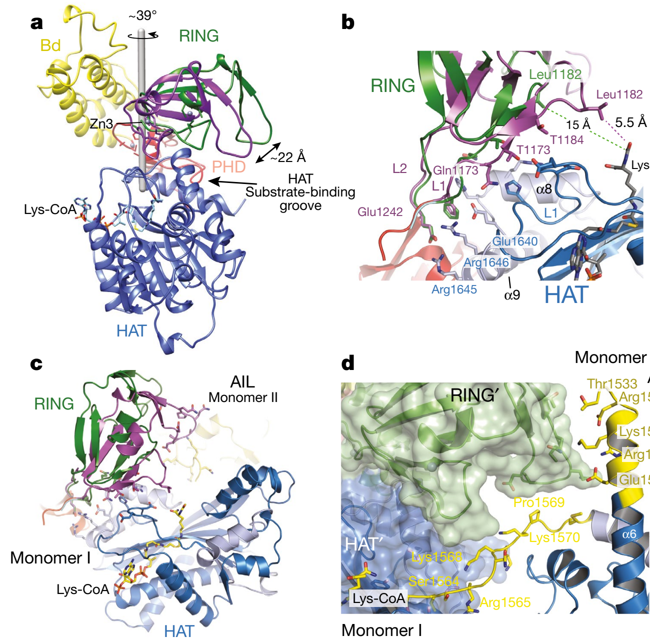


Fig. 3 | Structural rearrangement of the RING domain. **a**, The RING domain (green) rotates by approximately 39°, resulting in a displacement of about 22 Å away from the active site. The rotation axis is indicated as a grey rod. **b**, In the loop-swap conformation, residues in the RING–HAT interface are disrupted resulting in a more open HAT active site. Leu1182 is positioned approximately 15 Å away from the Lys-CoA inhibitor in the loop-swap conformation (green) but within 5.5 Å in the absence of the loop swap (magenta). **c**, Repositioning of the RING domain enables the AIL from monomer II to approach the active site of HAT on monomer I. **d**, Details of the interaction surface of the AIL from monomer II with the RING domain of monomer I.

domains from two neighbouring molecules are closely apposed (Fig. 2a). In all protomers, AIL residues 1520–1532 adopt a helical extension of α6 which packs against the outwardly rotated RING domain of the neighbouring protomer (Fig. 2a). In monomer II, residues 1566–1581 extend away from the HAT domain and associate with the substrate-binding pocket of the HAT domain in monomer I, approximately 17 Å away from the lysine substrate binding tunnel (Fig. 2b). The remainder of the AIL (residues 1532–1564) is disordered. In this conformation, positively charged residues K1568, K1569 and K1570 project towards the highly electronegative substrate-binding pocket of the HAT domain in monomer I (Fig. 2c). Analysis by size-exclusion chromatography–multi-angle laser-light scattering (SEC–MALS) revealed that p300 is monomeric at low micromolar concentrations (see Extended Data Fig. 6), which suggests that the AIL loop-swapped interactions do not appear to mediate the formation of stable dimers, but may instead constitute more transient self-associations. Although the AIL is clearly flexible and the electron density over the exchanged region is not visible in all protomers (Fig. 2b, c, Extended Data Fig. 2b, c), this arrangement supports the interpretation that, at high concentrations and when in close proximity to each other, two p300 monomers can engage each other through an AIL loop-swap.

Structural rearrangement of the RING domain

We previously proposed that active-site restriction by the RING domain is a negative regulatory mechanism for HAT activity[26]. A restricted active site is predicted to reduce the probability of substrates engaging with the active site by random diffusion, and could thus be important in enabling the regulation of acetylation by substrate recruitment. Consistent with this model, mutations that map to the structural framework that holds the RING domain in place result in HAT activation in cells[26]. In our current structure the RING domain rotates by around 39° away from the HAT active site, which results in an overall displacement by around 22 Å as compared to the previously determined structure that lacks the AIL (Fig. 3a). The axis of rotation is located perpendicular to the flexible loops L1 and L2 that connect the RING to the PHD domain.

The inward-rotated conformation (magenta in Fig. 3a) is stabilized by interactions between Glu1242 of the RING domain and Arg1645 and Arg1646 of helix α9 of the HAT domain. In addition, Gln1173, Thr1174 and Thr1184 of the RING domain pack against the unusually long loop (L1) in the HAT domain that covers the CoA portion of the Lys-CoA inhibitor. As a result, Leu1182 resides within about 5.5 Å of the lysine moiety of Lys-CoA (Fig. 3b). This inward conformation of the RING domain thus restricts substrate access to the HAT domain: the incoming AIL from the neighbouring p300 monomer II would clash with the RING domain in the inward conformation (Fig. 3c).

In the outward-rotated conformation, the interactions that attach the RING domain to the HAT domain are mostly disrupted (Fig. 3b). Leu1182 is positioned around 15 Å away from the substrate-binding site and the RING domain is cradled by the AIL extension of helix α6 of the neighbouring p300 molecule (monomer II residues 1524–1533; Fig. 3d). Despite shape complementarity, with a small buried surface area of about 320 $Å^2$, the interface is predominantly polar, which is uncharacteristic of a typical protein–protein interface. However, this interaction could help to stabilize an outward-rotated conformation of the RING domain and a more open active site of the HAT, apparently to enable access of the AIL and *trans*-acetylation.

Regulation of HAT activity by flanking domains

To systematically analyse the flanking domains, we generated a series of p300 constructs (Extended Data Fig. 4a) and analysed the effect on HAT activity in vitro and in cells. Overexpression of p300 generally resulted in hyperacetylated, active p300 variants (Extended Data Fig. 4b, c), which probably masks the functional role of structural elements potentially involved in autoinhibition of deacetylated p300. Deletion of the RING domain did not considerably alter autoacetylation or histone acetylation (Extended Data Fig. 5a). This deletion did

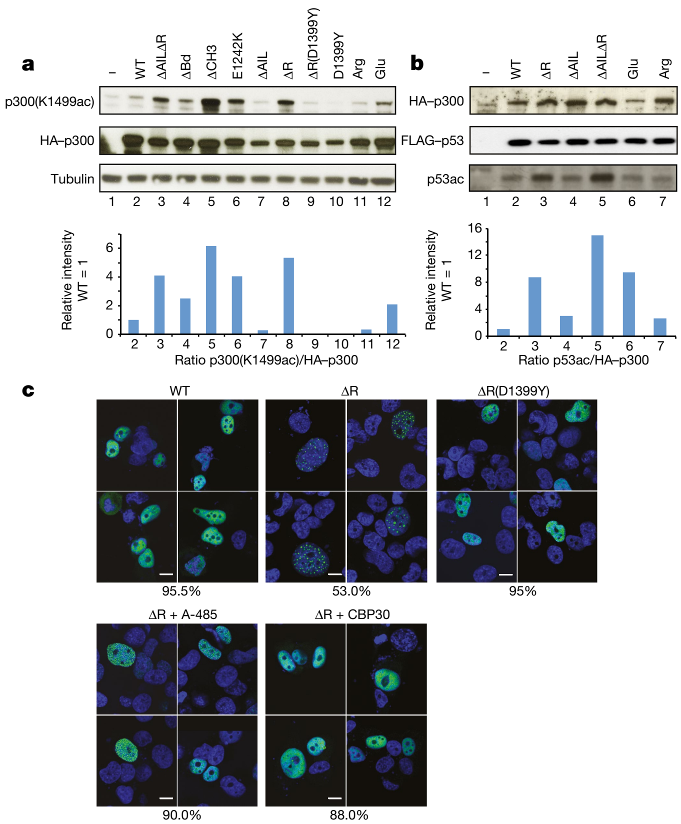


Fig. 4 | Regulation of HAT activity by flanking domains. a, Top, indicated variants of p300 were transiently co-transfected with p53 in COS cells and samples were analysed by western blotting using the indicated antibodies. Bottom, quantification of the p300(K1499ac) signal. WT, wild type. **b**, Top, analysis of p53 acetylation. Bottom, quantification of the p53 acetylation signal. Representative data of three independent experiments are shown. For details of the mutants, see Extended Data Fig. 4a. Arg and Glu indicate proteins in which lysine amino acids in the AIL segment spanning amino acids 1546–1570 were mutated to arginine or glutamate, respectively. **c**, Top, H1299 cells were transfected with the indicated constructs and analysed by immunofluorescence using anti-HA for p300 (green); cell nuclei were stained with Hoechst (blue). Bottom, cells were treated with the A-485 HAT or the CBP30 bromodomain inhibitor. The percentage of cells showing the indicated phenotype ($n = 200$ cells) is indicated below each panel. Scale bars, 10 μm. For gel source data, see Supplementary Fig. 1.

not adversely affect the structural integrity of p300, as shown by a crystal structure of the BΔRP module containing this deletion (Extended Data Fig. 5c).

Deletion of the AIL (ΔAIL) in all constructs resulted in decreased histone acetylation, but bromodomain deletion (ΔBd) did not affect HAT function (Extended Data Fig. 5a, b). Together, our results are consistent with previous observations of CBP that—at least in the active, hyperacetylated state of the enzyme—RING deletion does not substantially affect HAT activity and that the p300 AIL positively contributes to substrate acetylation[27]. We next introduced mutations into full-length p300 and monitored their effect on p300 autoacetylation and p53 acetylation upon transient co-overexpression in cells. Deletion of the RING (ΔR) and CH3 domains resulted in markedly increased p300 autoacetylation and p53 acetylation, but deletion of the bromodomain or AIL had no major effect (Extended Data Fig. 5e). As expected, introduction of the catalytic mutants D1399Y or Y1467F abolished p300 autoacetylation or p53 acetylation (Extended Data Fig. 5e). Immunofluorescence analysis showed that wild-type p300 as well as a ΔBd and ΔAIL deletion were uniformly distributed in the nucleus, but that the HAT-activating p300 variants ΔR and ΔCH3 formed nuclear foci that co-localized with p53 (Extended Data Fig. 5d).

To validate these results, we analysed and confirmed the phenotype of p300 mutants and p53 acetylation in another cell line (Fig. 4a, b). In addition, we analysed p300 variants in which eleven lysine amino acids (spanning amino acids 1546–1570 of the AIL) were mutated to arginine or glutamate, and found reduced or slightly increased p300 autoacetylation or p53 acetylation levels, respectively (Fig. 4a, b).

As we observed the formation of nuclear foci only with HAT-activating variants, we proposed that hyperacetylation drives p300 to form biomolecular condensates in cells. Accordingly, the introduction of a HAT-inactivating D1399Y mutation into p300 ΔRING, treatment with the p300 HAT inhibitor A-485[28], or treatment with the p300/CBP bromodomain inhibitor CBP30 greatly reduced foci formation (Fig. 4c). We therefore conclude that HAT activation drives biomolecular condensation of p300 in cells, apparently through substrate engagement by the bromodomain.

Regulation of p300 by the AIL and RING domain

We next sought to understand how the highly conserved and intrinsically disordered AIL segment contributes to the regulation of the catalytic function of p300. The AIL spans amino acid residues 1532–1567 and is positively charged in the deacetylated state, with an estimated isoelectric point (pI) of 10.9, and net charge of +7 at neutral pH. By contrast, upon autoacetylation of residues spanning Lys1542–1560[29], we estimate a pI of 3.5 and a net charge of −2. As the proximal substrate-binding groove of p300 is largely acidic (Fig. 2c), we proposed—consistent with earlier predictions[30]—that a deacetylated AIL would engage the substrate binding site through electrostatic interactions, presumably to prevent access of positively charged lysine-containing substrates. Given the disordered nature of the AIL, this proposed interaction is expected to be highly dynamic[27].

We tested this hypothesis through all-atom Monte Carlo simulations[31]. To make this approach tractable, our simulations held the backbone dihedral angles associated with the folded domains fixed, but all other degrees of freedom, including all backbone and side chain dihedral angles in the AIL, were fully sampled. As a result, these simulations should be seen to assess how the AIL interacts with the remainder of p300 given the observed crystal structure. Simulations were performed on the AIL in the deacetylated and acetylated states in the context of the p300 monomer. These simulations enabled us to investigate how acetylation influenced the conformation and intramolecular interactions of the AIL.

Simulations of the deacetylated AIL revealed the presence of extensive yet highly degenerate electrostatic interactions between the AIL and the RING domain and between the AIL and the HAT substrate-binding site. These interactions were quantifiable in terms of the normalized distances between pairs of amino acid residues (Fig. 5a, Supplementary Video 1). Lysine residues in the AIL dynamically associate through long-range electrostatic interactions with acidic residues (E1334, E1442, E1505, D1622, D1625 and D1628) in the p300 HAT substrate-binding pocket (Fig. 5c). The importance of these residues for substrate acetylation has been shown previously[32], and nuclear magnetic resonance data for CBP confirm that the AIL is intrinsically disordered in the deacetylated state[27].

By contrast, in the acetylated state we found no interactions between the AIL and the substrate binding site (Fig. 5b and Supplementary Video 2). The acetylated AIL essentially behaved like a self-avoiding random coil without any strong biases for interaction with itself or with the surrounding folded domains, including the bromodomain. It has been proposed that the AIL of CBP, when acetylated on K1596 (K1558 in p300), engages the bromodomain intramolecularly, thus competing with histone binding and negatively regulating substrate acetylation[27]. Isothermal calorimetry experiments showed the highest binding affinity for multiacetylated peptides, including the diacetylated histone peptides H3(K14ac/K18ac) and H4(K12ac/K16ac), generally following the pattern KacNNNKac (Extended Data Table 2). Monoacetylated peptides typically had weaker binding affinity. A crystal structure of the H4(K12ac/K16ac) peptide bound to BΔRP (Extended Data Fig. 5c)

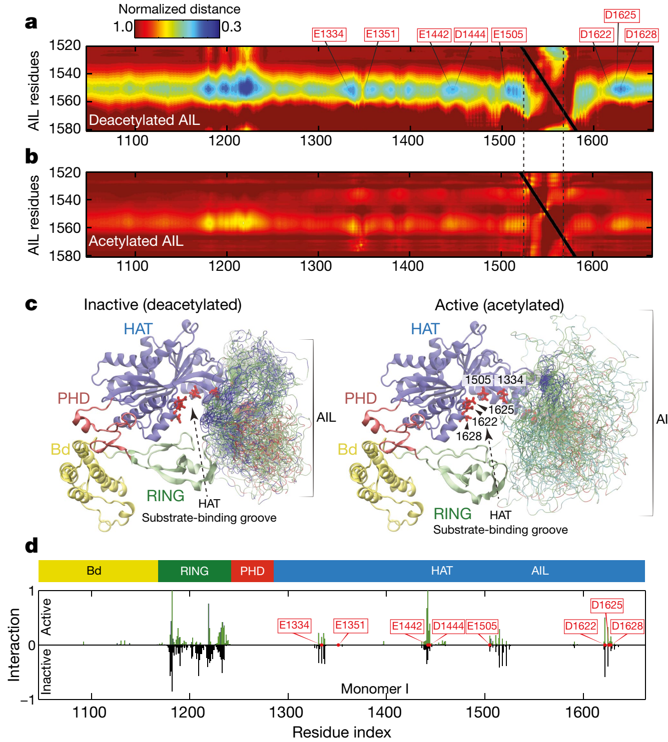


Fig. 5 | Acetylation of the AIL regulates dynamic interaction with the substrate-binding pocket of p300. a, Normalized distances between the AIL and residues in the inactive monomer. Inter-residue distances were normalized by the distances expected if the AIL behaved as a self-avoiding random coil. Electrostatic interaction mediated by conserved lysine residues between K1542 and K1560 of the AIL and aspartic acid or glutamic acid residues around the active site of the HAT domain (E1334, E1351, E1442, D1444, E1505, D1622, D1625 and D1628). The extensive contacts between the AIL and the RING domain originate in part from the proximity of the RING domain to the AIL in its inhibitory conformation. **b**, Normalized distances between the AIL and all residues in the active (acetylated) monomer. After acetylation, lysine-mediated electrostatic interactions are lost. **c**, Representative conformations with the AIL shown as an ensemble for the inactive deacetylated monomer (left) and the active acetylated monomer (right). The C_α atoms of residues in the AIL are coloured according to charge: blue, positive; red, negative; green, non-charged. The HAT substrate-binding groove is more exposed in the active acetylated state, due to both the relative position of the RING domain and the lack of preferential interactions by the AIL. **d**, Intermolecular interactions in the loop-swapped dimer between the AIL of one HAT and the adjacent subunit of the other. The adjacent subunit is either in the active (top) or inactive (bottom) conformation. In the active state, the AIL is able to directly engage with residues E1442 and E1444 from the adjacent HAT substrate-binding groove, which suggests that certain orientations of the RING domain can sterically hinder access to the AIL.

confirmed the acetyllysine-specific binding mode. However, a AIL peptide acetylated on the three lysines K1549, K1558 and K1560—corresponding to some of the most highly acetylated residues in the AIL[29]—failed to bind to the BRP module. Thus our interpretation is that the multiacetylated AIL is not a substrate for the bromodomain, presumably because of suboptimal spacing or sequence environment of the acetylated lysine sites of the AIL.

To understand how the RING domain influences the ability of substrates—including the AIL—to enter the active site of an adjacent p300 molecule, we performed simulations of the AIL in the context of the loop-swapped dimer, using a harmonic potential to maintain the AIL in the active site in order to assess potential intermolecular interactions (Fig. 5d, Extended Data Fig. 6a). In the active RING conformation, the AIL is able to engage the substrate binding site. However, in the inactive conformation, the frequency of contacts between the AIL and the acidic active site residues E1442 and D1444—residues proximal to the lysine substrate binding tunnel—was reduced by 70–75% (Fig. 5d). Thus, in the inactive conformation, the RING domain at least partially reduces catalytic activity by limiting accessibility of the active site to the AIL and other substrates.

One prediction from our models is that the deacetylated form of p300 adopts a more compact conformation, owing to dynamic engagement of the AIL with the HAT substrate-binding site, whereas the acetylated form adopts a more 'open' conformation (Fig. 5d). To test this possibility, we produced hypo- and hyperacetylated p300 variants (Extended Data Fig. 6e–g) and analysed the preparations by SEC–MALLS. All preparations were monomeric at the concentration tested (2 mg ml^{-1}) (Extended Data Fig. 6b–d, Extended Data Table 3). Hyperacetylation of p300 BRP-HAT resulted in a small decrease in the elution volume, which is indicative of a larger hydrodynamic radius (Extended Data Fig. 6b). A similar result was obtained upon comparison of hyper- and hypoacetylated BRP-HAT-CH3 (Extended Data Fig. 6c). By contrast, a variant that lacks the AIL showed no change in the elution volume upon hyperacetylation (Extended Data Fig. 6c). Our data are therefore consistent with the model that the catalytic p300 'core' adopts a compact conformation in the hypoacetylated state, with autoacetylation resulting in a more extended conformation.

Discussion

Our findings provide detailed mechanistic insights into how cellular signalling controls the activity of a chromatin regulator. We propose a multi-step process for p300 HAT activation and signal transmission to chromatin (Extended Data Fig. 7a–d). In the basal state, the deacetylated AIL is expected to maintain an overall positively charged environment in close proximity to the active site of the enzyme, thus preventing access of positively charged lysine-rich substrates. Direct access to the CoA-binding tunnel and autoacetylation of the AIL *in cis* appears to be prohibited, in part due to the positioning of the RING domain (Fig. 5d).

Cellular signalling initiates phosphorylation of transcription factors, such as IRF3 or STAT1, which results in their activation and dimerization. The activated, dimeric transcription factors are in their DNA-binding-competent conformation and can engage two molecules of p300 in the nucleus, thus increasing the likelihood of AIL disengagement from its inhibitory position *in cis* and of its capture *in trans* by a second p300 molecule. Association of two p300 molecules does not necessarily require precise stereospecific interactions between the structured domains, because acetylation at several lysines in the AIL indicates a series of possible conformations in such transiently associating dimers. We predict that regulated oligomerization uncouples recruitment from HAT activation, which could explain why not all p300/CBP recruitment events result in chromatin acetylation and gene activation[12–17,33].

It has been proposed that enhancer RNA interacts with the AIL to regulate CBP HAT activity[34]. We have attempted to reproduce these results using Klf6, one of the most potent enhancer RNAs reported[34]. We could not detect p300 HAT activation using up to equimolar amounts of Klf6 (Extended Data Fig. 7e, g). We note that, in a previous study, CBP was purified in buffer containing EDTA; this is detrimental to the structure of p300/CBP owing to the presence of multiple zinc-binding domains[35]. When unfolded by incubation with EDTA, CBP and p300 have a high tendency to aggregate and to form non-specific interactions[35]. Paradoxically, as the HAT domain is not affected, inclusion of EDTA can have an 'activating' effect in biochemical assays, apparently due to such non-specific aggregation (Extended Data Fig. 7f). The detrimental effects of EDTA on the structure and function of p300/CBP need to be taken into account in the interpretation of such data.

The ability of certain histone-modifying enzymes to bind to the post-translational modification (PTM) they generate has led to models in which such enzymes might propagate modified chromatin domains by a positive-feedback loop[36]. According to this view histone PTMs and other chromatin modifications form an additional,

DNA-sequence-independent layer of the genome, which is read out by enzymes that recognize these modifications to 'epigenetically' regulate genomic function[36]. An alternative view proposes that histone PTMs ultimately depend on DNA-sequence-dependent recruitment of chromatin modifiers, and so do not necessarily form an independent 'epigenetic' layer of the genome[8,37–39]. The controversy has arisen because it has been difficult to disentangle, for most chromatin regulators, the relative contributions of DNA targeting and histone PTM substrate engagement to the overall chromatin-modification reaction.

We show that regulation of p300 is linked to the activation and oligomerization status of transcription factor ligands, and therefore conclude that specificity for p300-mediated chromatin acetylation arises mainly through transcription-factor-mediated and DNA-sequence-dependent genome targeting. The next question is how the bromodomain contributes to p300 function. Although it is clear that the bromodomain can engage acetylated histone peptides and bind to hyperacetylated chromatin[26,40], deletion or mutation of the bromodomain has no apparent effect on substrate acetylation[26,41], has only minimal effects in a haematopoiesis model system[42], and bromodomain inhibition does not adversely affect genome targeting of CBP[43].

We favour a model in which DNA binding provides the lead anchoring mechanism: local hyperacetylation increases the binding valency by enabling bromodomain substrate engagement, which further helps to compartmentalize the biochemical reaction and contributes to signal maintenance[40]. p300 HAT-activating mutants form biomolecular condensates in cells when transiently overexpressed (Fig. 4c, Extended Data Fig. 5d). Treatment with a HAT inhibitor or bromodomain inhibitor greatly reduces the formation of condensates, which indicates that hyperacetylation and bromodomain–substrate engagement are critical in driving assembly. The formation of condensates, possibly through phase-separation, may provide a mechanism to enable signal integration on enhancers and transcriptional control[44]. It will be critical to disentangle cause–effect relationships of DNA targeting, chromatin modification and histone PTM substrate engagement of other chromatin regulators[45–47].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0621-1.

Received: 14 February 2018; Accepted: 15 August 2018;
Published online 15 October 2018.

1. Chen, Q., Sun, L. & Chen, Z. J. Regulation and function of the cGAS–STING pathway of cytosolic DNA sensing. *Nat. Immunol.* **17**, 1142–1149 (2016).
2. Panne, D., McWhirter, S. M., Maniatis, T. & Harrison, S. C. Interferon regulatory factor 3 is regulated by a dual phosphorylation-dependent switch. *J. Biol. Chem.* **282**, 22816–22822 (2007).
3. Zhao, B. et al. Structural basis for concerted recruitment and activation of IRF-3 by innate immune adaptor proteins. *Proc. Natl Acad. Sci. USA* **113**, E3403–E3412 (2016).
4. Parekh, B. S. & Maniatis, T. Virus infection leads to localized hyperacetylation of histones H3 and H4 at the IFN-β promoter. *Mol. Cell* **3**, 125–129 (1999).
5. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-beta enhanceosome. *Cell* **129**, 1111–1123 (2007).
6. Stark, G. R. & Darnell, J. E., Jr. The JAK-STAT pathway at twenty. *Immunity* **36**, 503–514 (2012).
7. Zhang, J. J. et al. Two contact regions between Stat1 and CBP/p300 in interferon gamma signaling. *Proc. Natl Acad. Sci. USA* **93**, 15092–15096 (1996).
8. Bedford, D. C. & Brindle, P. K. Is histone acetylation the most important physiological function for CBP and p300? *Aging* **4**, 247–255 (2012).
9. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
10. Visel, A. et al. ChIP–seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
11. Jin, Q. et al. Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. *EMBO J.* **30**, 249–262 (2011).
12. Bedford, D. C., Kasper, L. H., Fukuyama, T. & Brindle, P. K. Target gene context influences the transcriptional requirement for the KAT3 family of CBP and p300 histone acetyltransferases. *Epigenetics* **5**, 9–15 (2010).
13. Zhao, L. et al. Integrated genome-wide chromatin occupancy and expression analyses identify key myeloid pro-differentiation transcription factors repressed by Myb. *Nucleic Acids Res.* **39**, 4664–4679 (2011).
14. Waltzer, L. & Bienz, M. *Drosophila* CBP represses the transcription factor TCF to antagonize Wingless signalling. *Nature* **395**, 521–525 (1998).
15. Holmqvist, P. H. & Mannervik, M. Genomic occupancy of the transcriptional co-activators p300 and CBP. *Transcription* **4**, 18–23 (2013).
16. Kasper, L. H., Qu, C., Obenauer, J. C., McGoldrick, D. J. & Brindle, P. K. Genome-wide and single-cell analyses reveal a context dependent relationship between CBP recruitment and gene expression. *Nucleic Acids Res.* **42**, 11363–11382 (2014).
17. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
18. Thompson, P. R. et al. Regulation of the p300 HAT domain via a novel activation loop. *Nat. Struct. Mol. Biol.* **11**, 308–315 (2004).
19. Qin, B. Y. et al. Crystal structure of IRF-3 in complex with CBP. *Structure* **13**, 1269–1277 (2005).
20. Larabi, A. et al. Crystal structure and mechanism of activation of TANK-binding kinase 1. *Cell Rep.* **3**, 734–746 (2013).
21. Levy, D. E. & Darnell, J. E., Jr. Stats: transcriptional control and biological impact. *Nat. Rev. Mol. Cell Biol.* **3**, 651–662 (2002).
22. Shuai, K., Stark, G. R., Kerr, I. M. & Darnell, J. E., Jr. A single phosphotyrosine residue of Stat91 required for gene activation by interferon-gamma. *Science* **261**, 1744–1746 (1993).
23. Chen, X. et al. Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. *Cell* **93**, 827–839 (1998).
24. Wojciak, J. M., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. Structural basis for recruitment of CBP/p300 coactivators by STAT1 and STAT2 transactivation domains. *EMBO J.* **28**, 948–958 (2009).
25. Darnell, J. E., Jr. STATs and gene regulation. *Science* **277**, 1630–1635 (1997).
26. Delvecchio, M., Gaucher, J., Aguilar-Gurrieri, C., Ortega, E. & Panne, D. Structure of the p300 catalytic core and implications for chromatin targeting and HAT regulation. *Nat. Struct. Mol. Biol.* **20**, 1040–1046 (2013).
27. Park, S. et al. Role of the CBP catalytic core in intramolecular SUMOylation and control of histone H3 acetylation. *Proc. Natl Acad. Sci. USA* **114**, E5335–E5342 (2017).
28. Lasko, L. M. et al. Discovery of a selective catalytic p300/CBP inhibitor that targets lineage-specific tumours. *Nature* **550**, 128–132 (2017).
29. Karanam, B., Jiang, L., Wang, L., Kelleher, N. L. & Cole, P. A. Kinetic and mass spectrometric analysis of p300 histone acetyltransferase domain autoacetylation. *J. Biol. Chem.* **281**, 40292–40301 (2006).
30. Karanam, B. et al. Multiple roles for acetylation in the interaction of p300 HAT with ATF-2. *Biochemistry* **46**, 8207–8216 (2007).
31. Vitalis, A. & Pappu, R. V. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **30**, 673–699 (2009).
32. Liu, X. et al. The structural basis of protein acetylation by the p300/CBP transcriptional coactivator. *Nature* **451**, 846–850 (2008).
33. Soutoglou, E. et al. Transcription factor-dependent regulation of CBP and P/CAF histone acetyltransferase activity. *EMBO J.* **20**, 1984–1992 (2001).
34. Bose, D. A. et al. RNA binding to CBP stimulates histone acetylation and transcription. *Cell* **168**, 135–149.e122 (2017).
35. Matt, T., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. The CBP/p300 TAZ1 domain in its native state is not a binding partner of MDM2. *Biochem. J.* **381**, 685–691 (2004).
36. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17**, 487–500 (2016).
37. Ptashne, M. Epigenetics: core misconcept. *Proc. Natl Acad. Sci. USA* **110**, 7101–7103 (2013).
38. Rando, O. J. Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr. Opin. Genet. Dev.* **22**, 148–155 (2012).
39. Henikoff, S. & Shilatifard, A. Histone modification: cause or cog? *Trends Genet.* **27**, 389–396 (2011).
40. Nguyen, U. T. et al. Accelerated chromatin biochemistry using DNA-barcoded nucleosome libraries. *Nat. Methods* **11**, 834–840 (2014).
41. Rack, J. G. M. et al. The PHD finger of p300 influences its ability to acetylate histone and non-histone targets. *J. Mol. Biol.* **426**, 3960–3972 (2014).
42. Kimbrel, E. A. et al. Systematic in vivo structure–function analysis of p300 in hematopoiesis. *Blood* **114**, 4804–4812 (2009).
43. Picaud, S. et al. Generation of a selective small molecule inhibitor of the CBP/p300 bromodomain for leukemia therapy. *Cancer Res.* **75**, 5106–5119 (2015).
44. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169**, 13–23 (2017).
45. Coleman, R. T. & Struhl, G. Causal role for inheritance of H3K27me3 in maintaining the OFF state of a *Drosophila* HOX gene. *Science* **356**, eaai8236 (2017).
46. Laprell, F., Finkl, K. & Müller, J. Propagation of polycomb-repressed chromatin requires sequence-specific recruitment to DNA. *Science* **356**, 85–88 (2017).
47. Wang, X. & Moazed, D. DNA sequence-dependent epigenetic inheritance of gene silencing and histone H3K9 methylation. *Science* **356**, 88–91 (2017).

Acknowledgements This work was supported by grant 16-0280 from Worldwide Cancer Research. E.O. was supported by an EMBL Interdisciplinary Postdoctoral (EIPOD) fellowship. S.R. was supported by the Fondation ARC pour la recherche sur le Cancer and by the Fondation FINOVI. A.S.H. is a postdoctoral fellow in the laboratory of R.V. Pappu at Washington University in St. Louis. The computational work was supported by the Human Frontiers

Science Program (grant RGP0034/2017 to R.V. Pappu) and the St Jude Collaborative Research Consortium on Membraneless Organelles (to R.V. Pappu). We thank the staff at the European Synchrotron Radiation Facility (ESRF) beamlines ID29; L. Signor for mass spectroscopy analysis; R. Vance for the plasmid encoding GST-STING; and P. Cole for the A-485 inhibitor. S.K. and D.P. were supported by ANR Episperm3 program. S.K. received additional support from Fondation ARC Canc'air project (RAC16042CLA), Plan Cancer (CH7-INS15B66 and ASC16012CSA) and the Université Grenoble Alpes ANR-15-IDEX-02 LIFE and IDEX SYMER.

Reviewer information *Nature* thanks L. Chen, V. Hilser and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions E.O. designed and performed most experiments, analysed and validated the data and revised the draft with assistance from S.R., Z.I., N.H. and J.G. A.S.H. performed computational modelling and revised the draft. S.K. provided supervision, funding acquisition and commented on the draft. D.P. was involved in conceptualization, supervision, project administration, funding acquisition and wrote the original and revised drafts.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0621-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0621-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.P.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

Constructs. For cell-free protein expression, cDNA of p300 (NCBI reference sequence: NM_001429.3) variants were cloned into the pIVEX2.4d vector (Roche) with a N-terminal 6× His tag and a C-terminal Flag tag. In the ΔR constructs, the RING domain encompassing residues 1169–1241 was replaced by a glycine amino acid residue linker. In the ΔAIL constructs, loop amino acid residues comprising residues 1520–1581 were replaced by the flexible linker sequence SGGSG. For *Escherichia coli* expression, cDNA encoding residues 1048–1282, for the BRP or BΔRP were cloned into the vector pETM-33 (EMBL) with a tobacco etch virus (TEV)-cleavable N-terminal glutathione *S*-transferase (GST) tag. p300 BRP_HAT variants were cloned into pFASTBAC1 (Thermo Fisher) and expressed in insect cells as shown previously[26]. p300s constructs, spanning amino acid residues 324–2094, were cloned into pFASTBAC1 vector with an N-terminal Flag tag. Haemagglutinin (HA)-tagged full-length p300 variants were cloned into pcDNA3.1 (Thermo Fisher). Point mutations were introduced by QuikChange mutagenesis (Agilent). Point mutations and nucleotide deletions carried out in p300 full-length (1–2414) or p300s (324–2094) were done through transfer vectors as described previously[26]. STAT1ΔN (136–748), STAT1ΔNC (136–713) and IRF3ΔC (1–382) with a C-terminal intein tag were cloned into the pTXB1 vector (New England Biolabs) using the restriction enzymes NdeI (STAT1) or NcoI (IRF3) and SpeI. IRF3 (1–427) with an N-terminal His-tag cleavable by TEV protease was cloned using the restriction enzymes NcoI and XhoI into the vector pETM-11 (EMBL). All constructs were confirmed by DNA sequencing.

Expression and purification. Expression and purification of Flag-tagged p300s constructs was done as described previously[2]. This method enables purification of p300s variants that are already pre-acetylated. Expression and purification of p300 BRP_HAT and SIRT2 were performed as described previously[26]. TBK1 was expressed in insect cells and purified as described previously[20]. Cell-free protein synthesis was done in a 50 μl reaction volume. In brief, 10 μg ml^{-1} of His-p300 variants in pIVEX2.4d were added to a reaction mixture containing 1 mM amino acid mix, 0.8 mM rNTPs (guanosine-, uracil-, and cytidine-5′-triphosphate ribonucleotides), 1.2 mM adenosine-5′-triphosphate, 55 mM HEPES, pH 7.5, 68 μM folinic acid, 0.64 mM cyclic adenosine monophosphate, 3.4 mM dithiothreitol (DTT), 27.5 mM ammonium acetate, 2 mM spermidine, 5 μM $ZnCl_2$, 80 mM creatine phosphate, 208 mM potassium glutamate, 16 mM magnesium acetate, 250 μg ml^{-1} creatine kinase, 27 μg ml^{-1} T7 RNA polymerase, 0.175 μg ml^{-1} tRNA and 67 μl ml^{-1} S30 *E. coli* bacterial extract. Incubation was carried out at 22 °C with agitation for 16 h. Proteins were purified using Ni-NTA chromatography (IMAC Sepharose 6 FF, GE healthcare) in buffer 1 (20 mM TRIS, pH 8.0, 300 mM NaCl, 1 mM DTT, 5 μM $ZnCl_2$) containing Complete Protease Inhibitors EDTA-Free (Roche). The resin was washed with 20 column volumes of buffer 1 and the protein eluted with 5 column volumes of buffer 1 containing 300 mM Imidazole. The protein was concentrated in a pre-washed Amicon Ultra 0.5 ml Ultracel 10K centrifugal filter (molecular weight cut off 10 kDa; EMD Millipore). The protein was buffer-exchanged into buffer 1 using 0.5 ml Zeba Spin desalting columns (molecular weight cut off 7 kDa; Thermo Scientific), flash-frozen in liquid nitrogen and stored at −80 °C.

For expression of GST–BRP and GST–BΔRP fusion proteins in *E. coli* BL21 (DE3), LB medium enriched with 100 μM $ZnCl_2$ was used. Cell pellets were resuspended in buffer 1 containing Complete Protease Inhibitors EDTA-Free (Roche) and lysed using a microfluidizer (Microfluidics Corp.). The lysate was clarified by centrifugation for 30 min at 39,000*g* in a JA-25.5 rotor (Beckman) and applied to a Glutathione Sepharose 4 Fast Flow resin according to instructions from the manufacturer (GE Healthcare). The resin was washed with buffer 1 and incubated with His-tagged TEV protease (1:100 w/w) for 14–16 h at 4 °C. Subtractive Ni-NTA chromatography (IMAC Sepharose 6 FF, GE Healthcare) was then used to remove the residual His-tag and TEV protease. The untagged protein was further purified by gel filtration on a High Load 16/60 Superdex 75 column (GE Healthcare) equilibrated in 20 mM HEPES, pH 7.5, 300 mM NaCl, 0.5 mM TCEP and 5 μM $ZnCl_2$. The final protein was concentrated to 15 mg ml^{-1} in a prewashed Amicon Ultra-15 Centrifugal filter (molecular weight cut off 10 kDa; EMD Millipore), flash-frozen in liquid nitrogen and stored at −80 °C.

The expression and purification of non-phosphorylated STAT1 variants (STAT1ΔN, STAT1ΔNC) and IRF3ΔC (1–382) was done in *E. coli* using the IMPACT expression system (New England Biolabs). For the expression of Y701 phosphorylated variants (pSTAT1ΔN, pSTAT1ΔNC), proteins were co-expressed with Elk receptor tyrosine kinase domain in *E. coli* BL21(DE3) TKB1 cells (Agilent). Cells were collected by centrifugation and resuspended in buffer 2 (20 mM HEPES pH 7.5, 500 mM NaCl). The cells were lysed in a microfluidizer (Microfluidics Corp.) and the soluble fraction was obtained by centrifugation for 30 min at 39,000*g* in a JA-25.5 rotor (Beckman). The supernatant was first passed over chitin beads (New England Biolabs) and washed with buffer 2 for 10 column volumes. The protein was cleaved at 4 °C for 16 h in buffer 2 containing 50 mM DTT, eluted and further purified by gel filtration on a High Load 16/60 Superdex 200 column (GE Healthcare) equilibrated in buffer 2.

GST–STING, comprising the soluble cytoplasmic domain spanning amino acids 138–378, was expressed in *E. coli* BL21(DE3) at 37 °C for 3 h. The cells were collected by centrifugation and resuspended in buffer 3 (20 mM TRIS, pH 8.0, 300 mM NaCl, 1 mM DTT) containing Complete Protease Inhibitors EDTA-Free (Roche). The cells were lysed in a microfluidiser (Microfluidics Corp.) and the soluble fraction was obtained by centrifugation as above. The supernatant was passed over equilibrated Glutathione Sepharose 4 Fast Flow resin according to instructions by the manufacturer (GE Healthcare). The resin was washed with buffer 3 and eluted with 10 mM reduced Glutathione in buffer 3. The protein was further purified by gel filtration on a High Load 16/60 Superdex 200 column (GE Healthcare) equilibrated in 20 mM HEPES, pH 7.5, 300 mM NaCl, 0.5 mM TCEP. The final protein was concentrated to 16 mg ml^{-1} in a prewashed Amicon Ultra-15 centrifugal filter (molecular weight cut off 30 kDa; EMD Millipore), flash-frozen in liquid nitrogen and stored at −80 °C.

IRF3 was expressed in *E. coli* BL21(DE3) at 18 °C for 16 h. The cells were collected by centrifugation and resuspended in buffer 2 containing 10 mM imidazole. The cells were lysed in a microfluidiser (Microfluidics Corp.) and the soluble fraction was obtained by centrifugation as above. The supernatant was passed over Ni^{2+}-conjugated IMAC Sepharose resin (GE Healthcare) and washed with buffer 2 containing 20 mM imidazole. The protein was eluted in buffer 2 containing 500 mM imidazole and was further purified by gel filtration on a High Load 16/60 Superdex 200 column in buffer 2 containing 0.5 mM TCEP. IRF3 was phosphorylated in vitro at a 1:10 molar ratio TBK1:IRF3 (1 mg ml^{-1}) in presence of 5 mM $MgCl_2$ and 1 mM ATP. The reaction was incubated at 30 °C for 1 h and then for an additional 10 h at 21 °C. Phosphorylated IRF3 was further purified by size-exclusion chromatography on a Superdex S200 16/60 column (GE Healthcare) in 20 mM HEPES, pH 7.5, 300 mM NaCl, 0.5 mM TCEP. The production of recombinant histones was done following standard procedures[48].

Crystallization and structure determination. The p300 BRP_HAT construct comprising the AIL and the mutation Y1467F was deacetylated as described previously[26]. The protein at 4.5 mg ml^{-1} was incubated with a threefold molar excess of the bi-substrate inhibitor Lys-CoA[32] before crystallization. Crystals in the $P2_1$ space group were grown by hanging-drop vapour diffusion at 4 °C by mixing equal volumes of protein and crystallization solution containing 100 mM HEPES, pH 7.5, 18–22% polyethylene glycol 3350 and 0.2 M NaCl. Crystals were cryoprotected in 20–25% ethylene glycol and drop-frozen in liquid nitrogen. We collected native diffraction data to a minimum Bragg spacing of 3.1 Å resolution at the ESRF on beamline ID29 under a nitrogen gas stream at 100 K, at a wavelength of 1.282 Å. We processed the data with XDS (Extended Data Table 1). The structure of the p300 BRP_HAT was determined by molecular replacement using Phaser. There are four copies in the asymmetric unit and the RING domains were initially not visible in the electron density map and are partially disordered. Inspection of an anomalous difference map indicated peak density for the zinc ions and enabled positioning of the RING domain in the outward-rotated conformation. A final model was produced by iterative rounds of manual model building in Coot and refinement using PHENIX. The final model contains residues 1045 –1664 with a deletion of residues 1534–1567 and was refined to a 3.1 Å resolution with an R_{work} and an R_{free} of 19% and 26%, respectively (Extended Data Table 1). Analysis of the refined structure by MolProbity showed that there are no residues in disallowed regions of the Ramachandran plot. The MolProbity all-atom-clash score was 1.91, placing the structure in the 100th percentile among structures refined at 3.1 Å resolution ($n = 2,108$).

The BΔRP construct at 15 mg ml^{-1} was mixed with 2 mM of a 11-mer histone peptide H4 (10–20) GLGKacGGAKacRHR (only the underlined amino acid sequence is visible in the electron density map) containing two acetylated lysine residues at K12 and K16, H4(K12ac/K16ac). Crystals in the $P2_12_12_1$ space group were grown by hanging-drop vapour diffusion at 21 °C by mixing equal volumes of protein and crystallization solution containing 1.6 M ammonium sulfate and 100 mM bicine at pH 9.0. Crystals were cryoprotected in 20% ethylene glycol and drop-frozen in liquid nitrogen. We collected native diffraction data to a minimum Bragg spacing of 2.5 Å resolution at the ESRF on beamline ID29 under a nitrogen gas stream at 100 K, at a wavelength of 1.0 Å (Extended Data Table 1). Data processing, molecular replacement and refinement were performed as indicated above. The final model contains two copies of the BΔRP module corresponding to residues 1049–1279 of p300 in the asymmetric unit. As expected, replacement of the RING domain residues 1169–1241 by a single glycine amino acid linker did not adversely affect the remainder of the BP module. Analysis of the refined structure by MolProbity showed that there are no residues in disallowed regions of the Ramachandran plot. The MolProbity all-atom-clash score was 0.97, placing the structure in the 100th percentile ($n = 6,960$).

Monte Carlo simulations. All-atom Monte Carlo simulations were performed using ABSINTH implicit solvent model and version 2 of the CAMPARI Monte Carlo simulation engine (http://campari.sourceforge.net)[31]. The initial AIL loop was constructed using MODELLER, and the complete set of backbone and side chain torsional angles were sampled for the AIL for which electron density was missing. Simulation analysis was performed with MDTraj and CTraj (http://pappulab.wustl.edu/CTraj.html)[49]. The backbone degrees of freedom of the folded domains were not sampled, while all amino acid side chains were fully sampled. CAMPARI simulations explore conformational space through perturbation to the torsional angles (as opposed to Cartesian positions, as is typical for molecular dynamics). Consequently, a fully closed loop represents a major sampling challenge. To address this, we severed the covalent backbone bond between the N-terminal part of the AIL loop and the folded domain, and replaced this bond with a strong harmonic potential that recapitulates the distances constraint associated with the covalent bond. This allows moves to fully rotate the chain and markedly improves the efficiency of conformational sampling.

We generated 5,000 independent non-overlapping starting configurations and used a clustering approach to identify the most distinct 200 conformations. These were used as the starting conformations for full simulations. We ran 200 independent simulations of the deacetylated and acetylated p300 in the monomeric form, and 200 independent simulations of the loop-swapped p300 dimer in the active and inactive form (800 simulations total). Analysis was performed after an initial equilibration. Dimer simulations applied a harmonic potential between residue 1550 from the AIL and residue 1442 from the other monomer to maintain the AIL in the active site. This enabled us to directly compare active-site accessibility of the AIL. For monomer simulations, no restraints were applied.

Each residue on the folded structure was evaluated for contacts with any residue in the AIL, and these contacts were summed to give an effective contact score. In this manner, the residues on the folded structure that most frequently interacted with any residue on the AIL were directly identified. Interaction was primarily electrostatic in nature, with residues E1334, E1444, E1505, D1622, D1625 and D1628 engaging in direct interactions. There are also extensive interactions between the AIL and the RING domain, although we cannot rule out that these interactions are driven by the harmonic potential applied to pull E1442 towards the active site. As might be expected, the AIL–RING interactions differed between the active and inactive conformations.

To assess interactions between the AIL and the folded domains in the monomer simulations, scaling map analysis was performed. In this analysis, a simulation of the AIL as a true self-avoid random coil is performed to generate a reference state, and then the mean inter-residue distances obtained in the full simulations are normalized by the distances obtained from this reference. The self-avoiding random coil simulations are performed using an identical protocol to the full simulations, with the exception that the only contribution to inter-atomic interactions comes from the repulsive part of the Lennard-Jones potential, meaning no attractive inter-atomic interactions or solvation effects are experienced. This ensures we generate a sequence and structure-specific self-avoiding random coil ensemble that provides a true reference state. Extensive details on the technical aspects associated with the generation of this reference state have been described previously[50]. The scaling maps enable us to easily identify local regions that engage in interactions that cause deviations from self-avoiding random coil behaviour.

HAT assays. The standard autoacetylation HAT assay was performed using [^{14}C] acetyl-CoA (Perkin-Elmer). Autoacetylation of p300 was quantified by autoradiography after SDS–PAGE gel analysis. The p300s preparations were equilibrated in 1× HAT buffer (25 mM TRIS-HCl, pH 7.5, 100 mM NaCl, 1 mM DTT, 10% glycerol and 1× Complete EDTA-free protease inhibitor (Roche)) for 10 min at 30 °C before initiation of the reaction by the addition of 200 μM [^{14}C]acetyl-CoA for the indicated time points. For experiments containing IRF3 STAT1 or the enhancer RNA (eRNA) Klf6, autoacetylation assays were performed at a fixed equimolar concentration (2 μM) of p300s and the indicated transcription factor or Klf6. Assays were performed in triplicate with different batches of proteins and on different days. At the indicated time point, 5 μl of the reaction was quenched by addition of 5 μl of 2× SDS gel loading buffer followed by analysis on a 4–20% SDS–PAGE gel. Experiments shown in Fig. 1c were performed in 1× kinase buffer (20 mM HEPES, pH 7.5, 250 mM NaCl, 20 mM β-glycerol phosphate, 1 mM sodium vanadate, 10 mM $MgCl_2$, 1 mM DTT, 1 mM ATP and a mix of 20 μM [^{14}C]acetyl-CoA and 80 μM cold acetyl-CoA (A2056, Sigma). 1 μM of p300s was incubated in the presence or absence of 1 μM IRF3, 2 μM TBK1 and 1 μM STING (lanes 1–6) or increasing amounts of STING as indicated (lanes 7–12). The gels were analysed by western blotting as indicated below or were fixed for 30 min in a solution containing 3% glycerol, 10% glacial acetic acid, 20% ethanol (v/v/v) in water. The gels were soaked for 5 min in a solution containing 1% glycerol, 5% PEG8000 in water and were dried for 30 min using a Bio-Rad Gel Dryer and the radioactivity quantified on a phosphorimage analyser (Typhoon, GE Healthcare) followed by analysis using ImageJ 1.8.0_112[51].

A p300 HAT scintillation proximity assay was designed similar to that described previously[28]. In brief, as a substrate we used a synthetic histone H4 peptide containing 15 amino acids derived from the N terminus of human H4 that was chemically attached to biotin with an amino hexanoic linker (Biotin-C6- GRGKGGKGLGKGGAK) (from peptid.de). The synthetic peptide was re-suspended in water and adjusted to pH 7.0 with concentrated NaOH.

A typical reaction contained p300s (50 nM), 12.5 μM biotinylated H4 peptide, acetyl-CoA (0.1 μM to 10 μM set at around 10 times the apparent K_m) in 20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 5 μM $ZnCl_2$, 0.01% Tween-20 and 0.1% BSA (w/v). For reactions containing pIRF3, 50 nM was added. 20 μl of a 2× reaction mixture containing p300s, H4 peptide with and without pIRF3 was pre-incubated at 30 °C for 5 min. The reaction was initiated by the addition of 20 μl of 2× acetyl-CoA containing a 1:3 mix of tritiated [^{3}H]acetyl-CoA (PerkinElmer; NET290050UC) with cold acetyl-CoA. For example, for 10 μM final acetyl-CoA concentration, a mix of 5 μM [^{3}H]acetyl-CoA and 15 μM cold acetyl-CoA (A2056, Sigma) was used. The reaction was quenched at the indicated time points by delivering 40 μl of the reaction mix into 120 μl of 0.5 M HCl in a FlashPlate Plus Streptavidin 96-well scintillant-coated microplate (Perkin Elmer, SMP103001PK). The plate was incubated for 1 h, and light emission was counted in a MicroBeta2 Scintillation Counter (Perkin Elmer) at 1 min per well in the top count mode. Counts per minute (cpm) were plotted against acetyl-CoA concentration. Typical progress curves are shown in Extended Data Fig. 1d. The initial rate was estimated by linear regression during the first 10 min of the reaction and plotted against acetyl-CoA concentration. All data were analysed using GraphPad Prism 7.0.

For the results shown in Fig. 4, acetylation reactions were performed in acetylation reaction buffer HAT (25 mM Tris-HCl, pH 7.5, 100 mM NaCl, 0.1 mM EDTA, 1 mM DTT, 10% glycerol and 1× complete EDTA-free protease inhibitor (Roche)) with 50 μM acetyl CoA (Sigma), 100 ng ml^{-1} trichostatin A and 2 μg of purified histone octamer. Reactions were incubated for 30 min at 30 °C and stopped by the addition of 3× SDS gel loading buffer, then used for Coomassie staining and immunoblotting.

Multi angle laser light scattering-size exclusion chromatography. Before SEC–MALLS runs, p300 variants were acetylated and deacetylated using p300 HAT or SIRT2 as described previously[26]. The reactions were analysed by liquid chromatography–mass spectrometry as described previously[52]. Size-exclusion chromatography was performed at a flow rate of 0.5 ml min^{-1} on a Superdex 200 Increase 10/300 GL column equilibrated in SEC–MALLS buffer (20 mM HEPES, 300 mM NaCl, 5 μM $ZnCl_2$, 0.5 mM TCEP) at 21 °C. A 50 μl sample of p300 at 2 mg ml^{-1} was injected onto the column and multi angle laser light scattering was recorded with a laser emitting at 690 nm using a DAWN-EOS detector (Wyatt Technology Corp.). The refractive index was measured using a RI2000 detector (Schambeck SFD). The molecular weight was calculated from differential refractive index measurements across the centre of the elution peaks using the Debye model for protein using ASTRA software version 6.0.5.3.

In vitro eRNA transcription. eKlf6 eRNA corresponding to 496 nucleotides of the sense strand of human chr13:5802100-5802596 (ref. [34]) was produced by in vitro transcription from a pMA plasmid containing a eKlf6 insert synthesized by GeneArt Gene Synthesis (Thermo Fisher). pMA_Klf6 plasmid (50 μg) was linearized with 80 U of KpnI-HF in a final volume of 100 μl and incubated at 37 °C for 14–16 h. The in vitro transcription reaction was done in a final volume of 1 ml, using 1× T7 buffer, T7 RNA Polymerase and 1 U of RNaseOUT Recombinant Ribonuclease Inhibitor (Thermo Fisher). After incubation for 2 h at 37 °C, 0.5 U of TURBO DNase (2 U μl^{-1}) (Thermo Fisher) and 1 μM $CaCl_2$ was added to the reaction and incubated for 30 min at 37 °C. Following DNaseI treatment, 2 μl of a 30 mg ml^{-1} stock of proteinase K powder (Thermo Fisher), dissolved in proteinase K buffer (10 mM TRIS pH 7.5, 1 mM $CaCl_2$ and 40% glycerol), was added and incubated for 45 min at 37 °C. Buffer was exchanged into 20 mM HEPES, pH 7.5, 300 mM NaCl, 0.5 mM TCEP using Amicon Ultra-0.5 ml centrifugal filters (molecular weight cut off 3 kDa, EMD Millipore). To further purify the RNA, 3 volumes of TRIzol (Thermo Fisher) was added to the RNA sample, followed by isopropanol precipitation. Purified Kfl6 RNA was resuspended in 20 mM HEPES, pH 7.5, 300 mM NaCl and 0.5mM TCEP. RNA was quantified using a Nanodrop spectrophotometer (Thermo Fisher). The quality of Klf6 was assessed by agarose gel electrophoresis in 1× TBE buffer or using denaturing 6M urea 14% PAGE (Extended Data Fig. 7c).

Immunoblotting, immunofluorescence and antibodies. We have used the cell lines COS and H1299. They are not on the list of commonly misidentified cell lines maintained by the International Cell Line Authentication Committee. COS cells were purchased from ATCC, product reference ATCC-CR1-1651, lot no. 4171903. H1299 cells were authenticated on December 6th 2016 by LGC Standards: Cell Line Authentication Service. Cell lines were authenticated using short tandem repeat (STR) analysis as described in 2012 in ANSI Standard (ASN-0002) Human cell lines were authenticated using standardization of STR Profiling by the ATCC Standards Development Organization as described in ref. [53]. Upon the receipt of COS cells

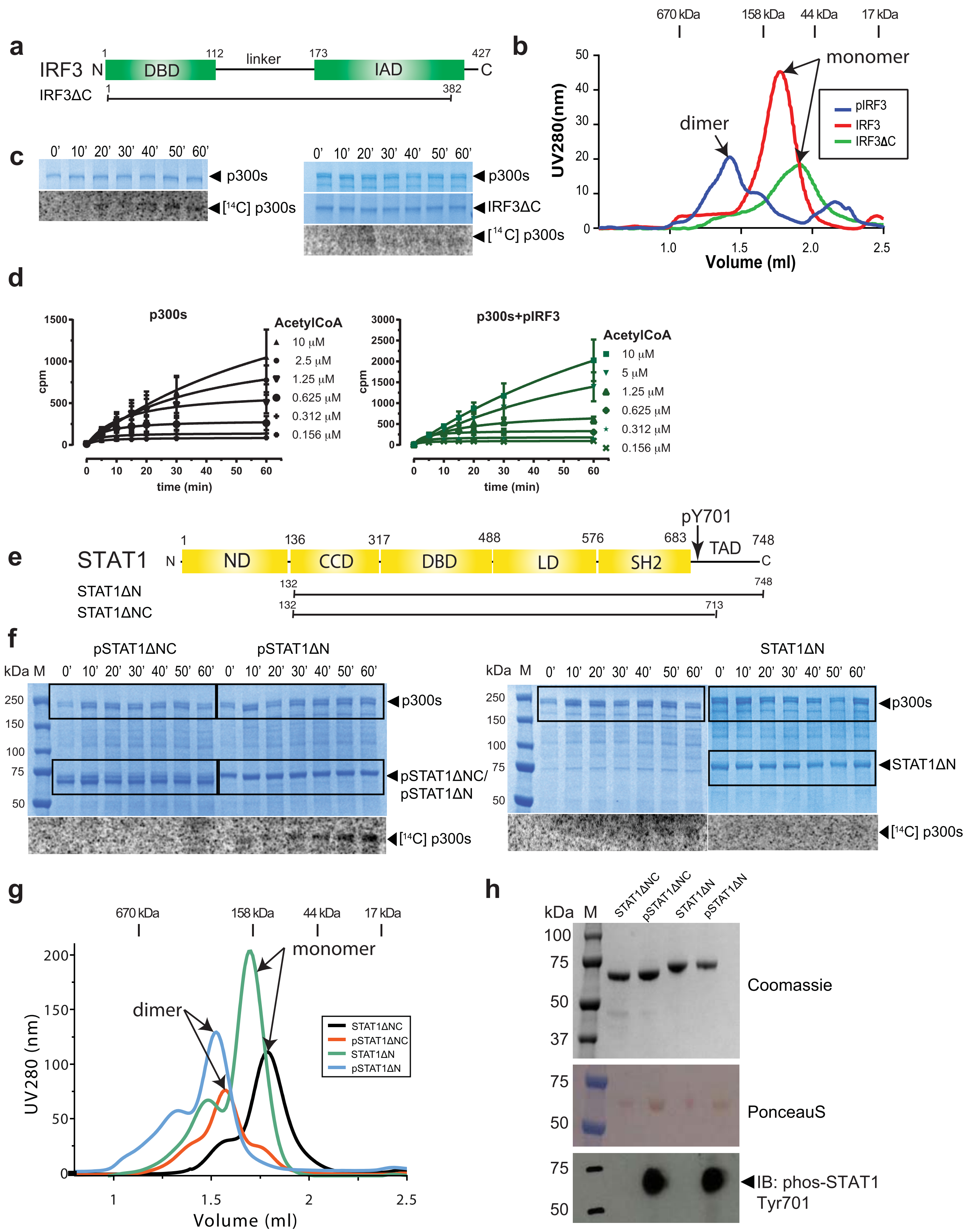
and after the authentication of H1299 cells, they were expanded. A mycoplasma contamination test (MycoAlert Mycoplasma Detection Kit, Lonza cat no. LT07-418) was performed and the mycoplasma-free cells were frozen and kept in liquid nitrogen. After thawing they were kept in culture for 30 passages with a mycoplasma contamination test after 15 passages. For immunoblotting, proteins were separated on 4–12% Bis-Tris SDS–PAGE gel (NuPAGE precast gel, Thermo Fisher) and transferred onto a nitrocellulose membrane (Hybond C+, GE Healthcare). Membranes were blocked with 5% skim milk in PBST buffer (PBS, 0.1% Tween-20) and probed with anti-p300(K1499ac) rabbit polyclonal antibody (1:2,500 dilution; Cell Signaling, 4771), anti-Kac rabbit polyclonal antibody (1:2,500 dilution; Cell Signaling, 9441), anti-Flag mouse monoclonal antibody (1:2,500 dilution; Sigma, F1804), anti-HA rabbit polyclonal antibody (1:2,500 dilution, Abcam, ab9110). For the detection of STAT1 or IRF3 phosphorylation, the membrane was blocked with 5% milk in PBST followed by overnight incubation at 4 °C with anti-phospho-Stat1 (Tyr701) rabbit monoclonal antibody (1:2,500 dilution; Cell Signaling no. 9171) in PBST buffer containing 5% BSA. For detection of IRF3 S396 phosphorylation, we used anti-phospho IRF3 S396 rabbit monoclonal antibody (1:2.500 dilution; Cell Signaling no. 4947). Incubations were performed as above. Membranes were washed extensively in PBST buffer before and after incubation with anti-rabbit or anti-mouse HRP-conjugated secondary antibody (1:10,000 dilution; GE Healthcare, NA934 or NA931), and protein bands were visualized on film after the ECL reaction (ECL Prime, GE Healthcare). Immunofluorescence was done as described previously[26] or as follows: 24 h post-transfection, cells were treated with a DMSO control or with 2.5 μM (final concentration) of HAT inhibitor A-485 (P. Cole, Harvard Medical School) or p300/CBP Bromodomain inhibitor (CBP30, Sigma, no. SML1133) dissolved in DMSO. The final DMSO concentration in all assays was 0.25%. After 24 h, cells were rinsed once with RNase-free PBS 1X. Next, cells were permeabilized in freshly prepared 0.2% Triton (Sigma) buffer for 5 min and were then fixed in freshly prepared 4% formalin solution (Sigma) for 10 min at room temperature. After three washes in RNase-free 1X PBS at room temperature, the cells were incubated in 5% skim milk in 1X PBS for 30 min and then probed with anti-HA high-affinity monoclonal antibody (1:100 dilution; Roche Applied Science, cat. no. 11867423001) overnight at 4 °C. Cells were washed extensively with 1X PBS before and after incubation with Alexa Fluor 488-conjugated secondary antibody (1:500 dilution; Invitrogen, no. A-11006) for 1 h at 37 °C. Cells were counterstained with Hoechst (250 ng ml^{-1}) and examined under a confocal laser scanning microscope (LSM510, Zeiss).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

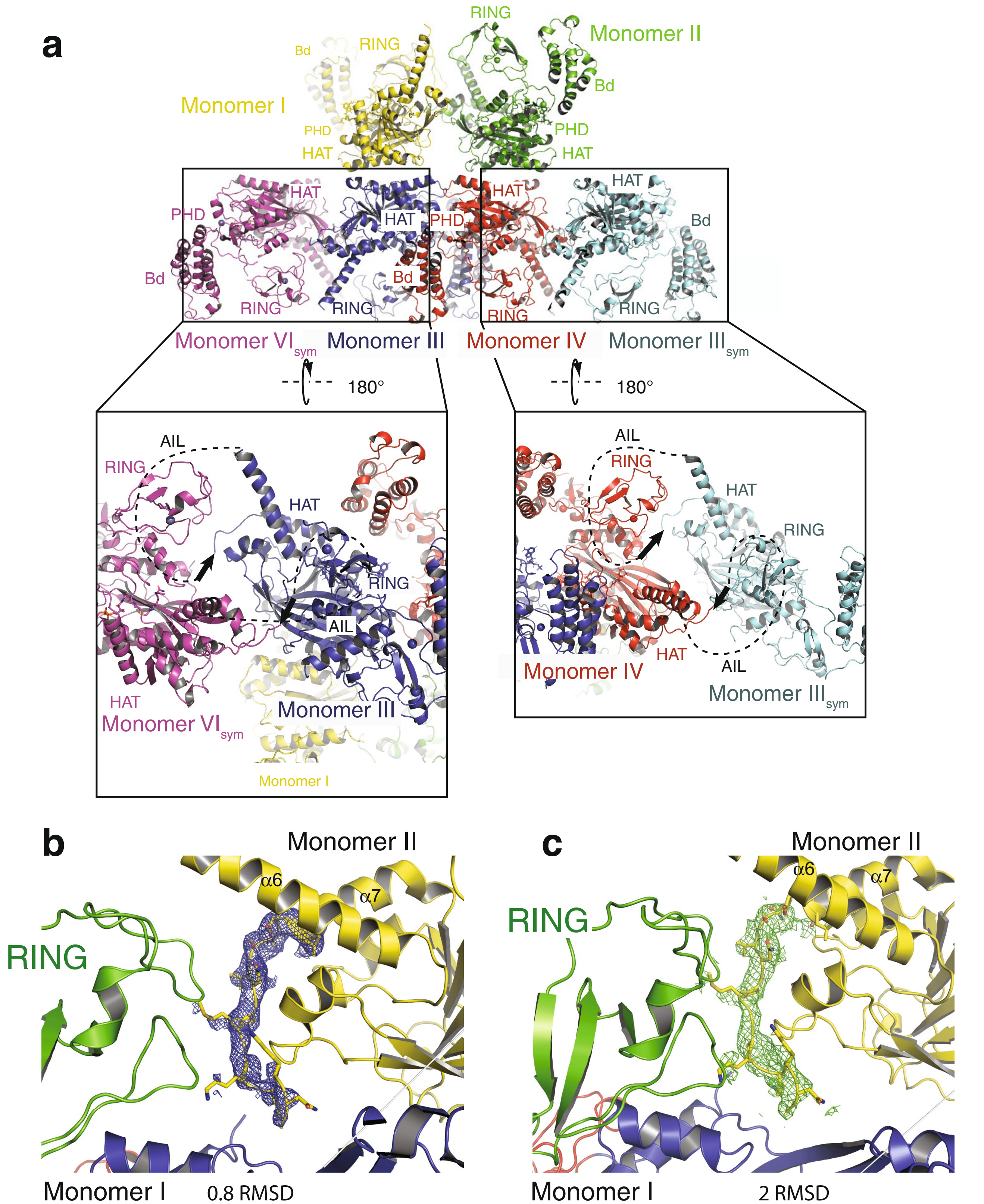
Coordinates for the p300 core structure and BΔRP bound to a diacetylated histone H4 peptide are available from the Protein Data Bank (PDB) under accession numbers 6GYR and 6GYT, respectively. Source data are available for Fig. 1b, f and Extended Data Fig. 1d. Figure 1d shows the initial velocities from reactions shown in Extended Data Fig. 1d.

48. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol*. **304**, 3–19 (1999).
49. McGibbon, R. T. et al. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J*. **109**, 1528–1532 (2015).
50. Holehouse, A. S., Garai, K., Lyle, N., Vitalis, A. & Pappu, R. V. Quantitative assessments of the distinct contributions of polypeptide backbone amides versus side chain groups to chain expansion via chemical denaturation. *J. Am. Chem. Soc*. **137**, 2984–2995 (2015).
51. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9**, 671–675 (2012).
52. Kaczmarska, Z. et al. Structure of p300 in complex with acyl-CoA variants. *Nat. Chem. Biol*. **13**, 21–29 (2017).
53. Capes-Davis, A. et al. Match criteria for human cell line authentication: where do we draw the line? *Int. J. Cancer* **132**, 2510–2519 (2013).
54. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res*. **14**, 1188–1190 (2004).

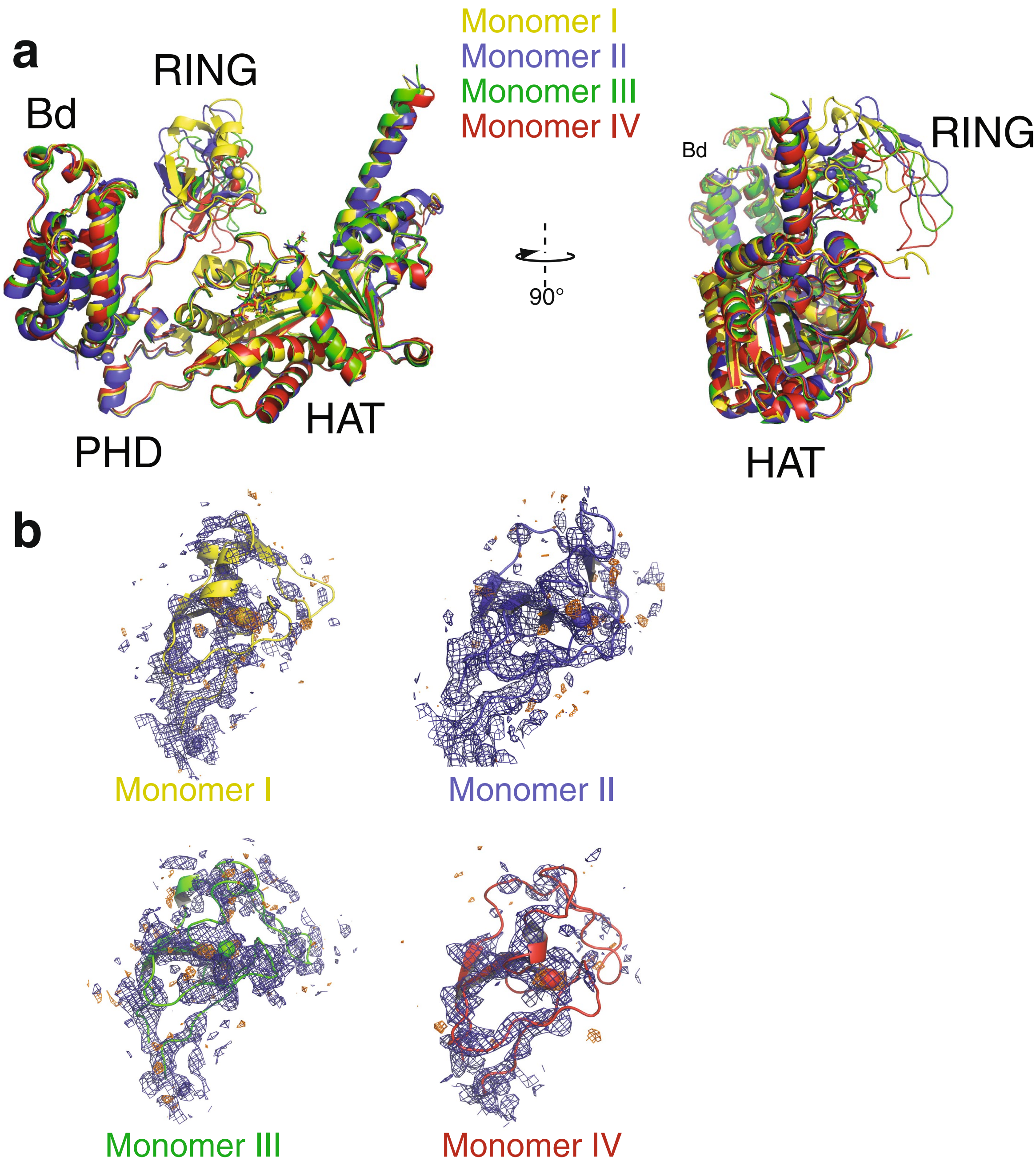


Extended Data Fig. 1 | See next page for caption.

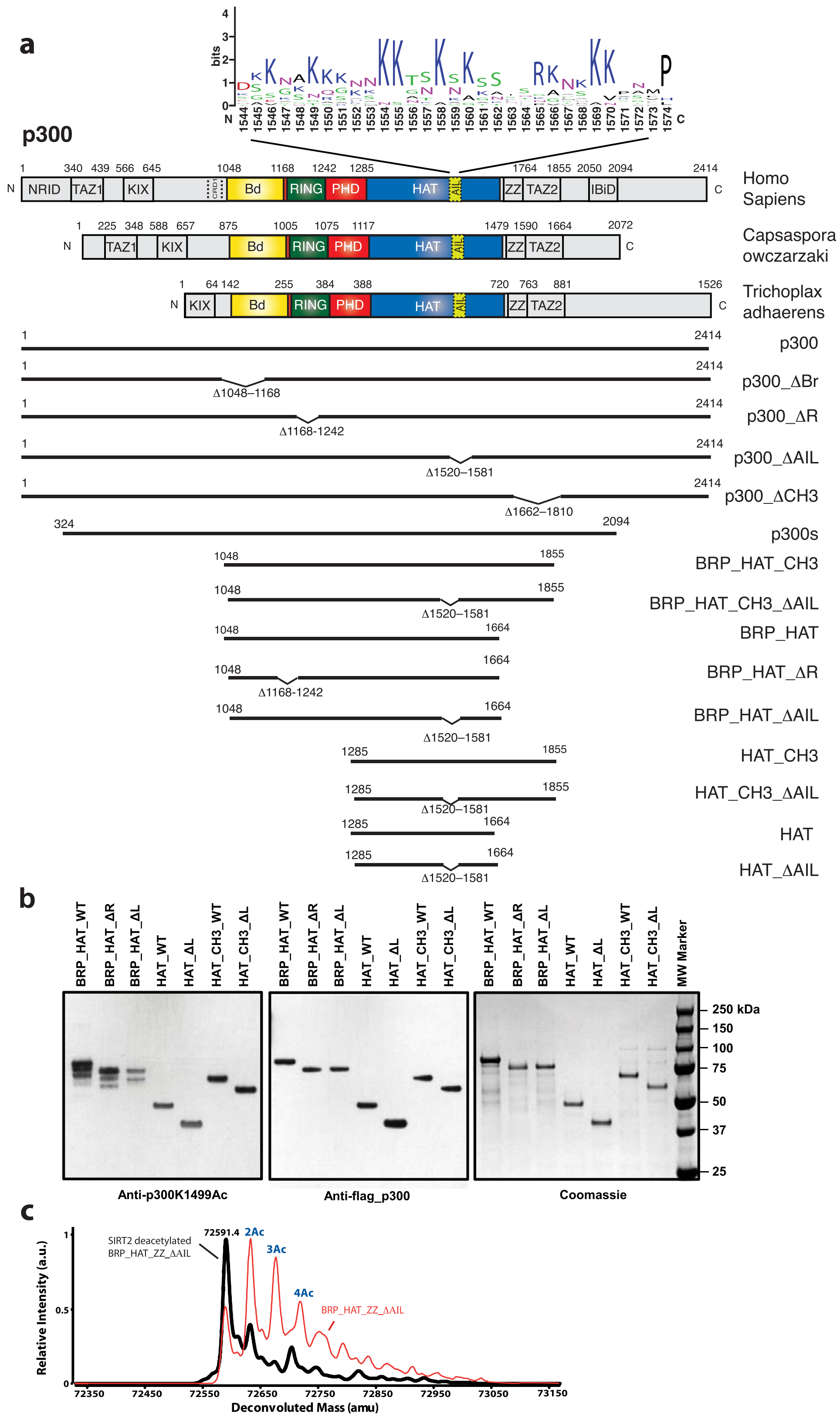
Extended Data Fig. 1 | The effect of IRF3 or STAT1 activation and oligomerization on p300 autoacetylation. a, The domain structure of IRF3. The truncation construct used is shown at the bottom. **b**, Size-exclusion chromatography of IRF3 variants. Red, unphosphorylated IRF3; blue, phosphorylated pIRF3; green, C-terminally truncated IRF3ΔC. Representative data of three independent experiments are shown. **c**, A constant amount of p300s (2 μM) was incubated alone or in the presence of C-terminally truncated IRF3ΔC (2 μM) for the indicated time points. Samples were analysed by SDS–PAGE followed by Coomassie staining and autoradiography. **d**, Progress curves of HAT scintillation proximity assay. Histone H4 substrate acetylation in the presence (green) or absence (black) of pIRF3 and varying concentrations of [^{3}H]acetyl-CoA. The degree of histone H4 substrate acetylation at different time points and the initial velocity (cpm min^{-1}) at the indicated acetyl-CoA concentrations were determined and plotted in Fig. 1e. Three independent experiments were performed and the mean value and error bars representing the standard deviation are shown. **e**, The domain structure of STAT1. The truncation constructs used are shown at the bottom, and the Tyr701 phosphorylation site is indicated. **f**, Uncropped images of SDS–PAGE gels shown in Fig. 1d. The ^{14}C autoacetylation signal of p300s is shown at the bottom. **g**, Size-exclusion chromatography of STAT1 variants. Black, STAT1ΔNC; green, STAT1ΔN; red, Y701-phosphorylated pSTAT1ΔNC; blue, Y701-phosphorylated pSTAT1ΔN. **h**, SDS–PAGE analysis of STAT1 variants and analysis by western blotting. Top, Coomassie staining of SDS–PAGE gel; middle, PonceauS staining; bottom, western blot using anti-Phospho-Stat1 (Tyr701). Representative data of three independent experiments are shown. For gel source data, see Supplementary Fig. 1.



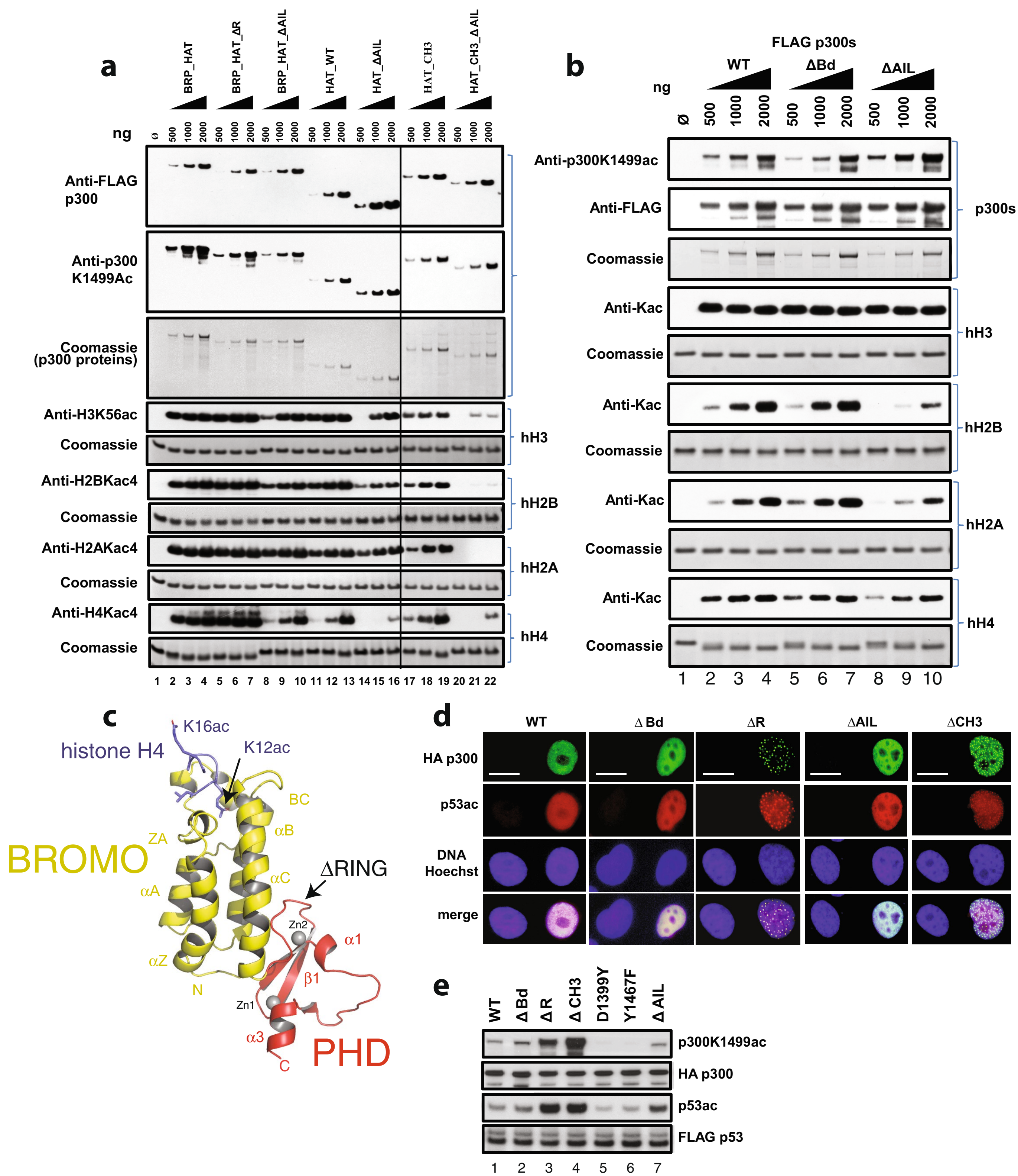
Extended Data Fig. 2 | Crystal packing of the p300 core molecule. **a**, There are four p300 molecules (monomers I–IV) in the asymmetric crystallographic unit. The four molecules show an antiparallel arrangement of the BRP-HAT domains. As a result, HAT domains from monomers I and II are closely apposed. Monomers III and IV engage monomer IV$_{sym}$ and monomer III$_{sym}$, respectively, of a neighbouring crystallographic unit, showing that all promoters are in a AIL loop-swap conformation. Black arrows indicate the direction of the AIL. The disordered segment of the AIL is shown as a black dotted line. **b**, **c**, Electron density of the AIL. $2F_o - F_c$ (**b**) and $F_o - F_c$ (**c**) difference density omit maps contoured at 0.8 and 2.0 r.m.s.d., respectively. Coloured as in Fig. 3.



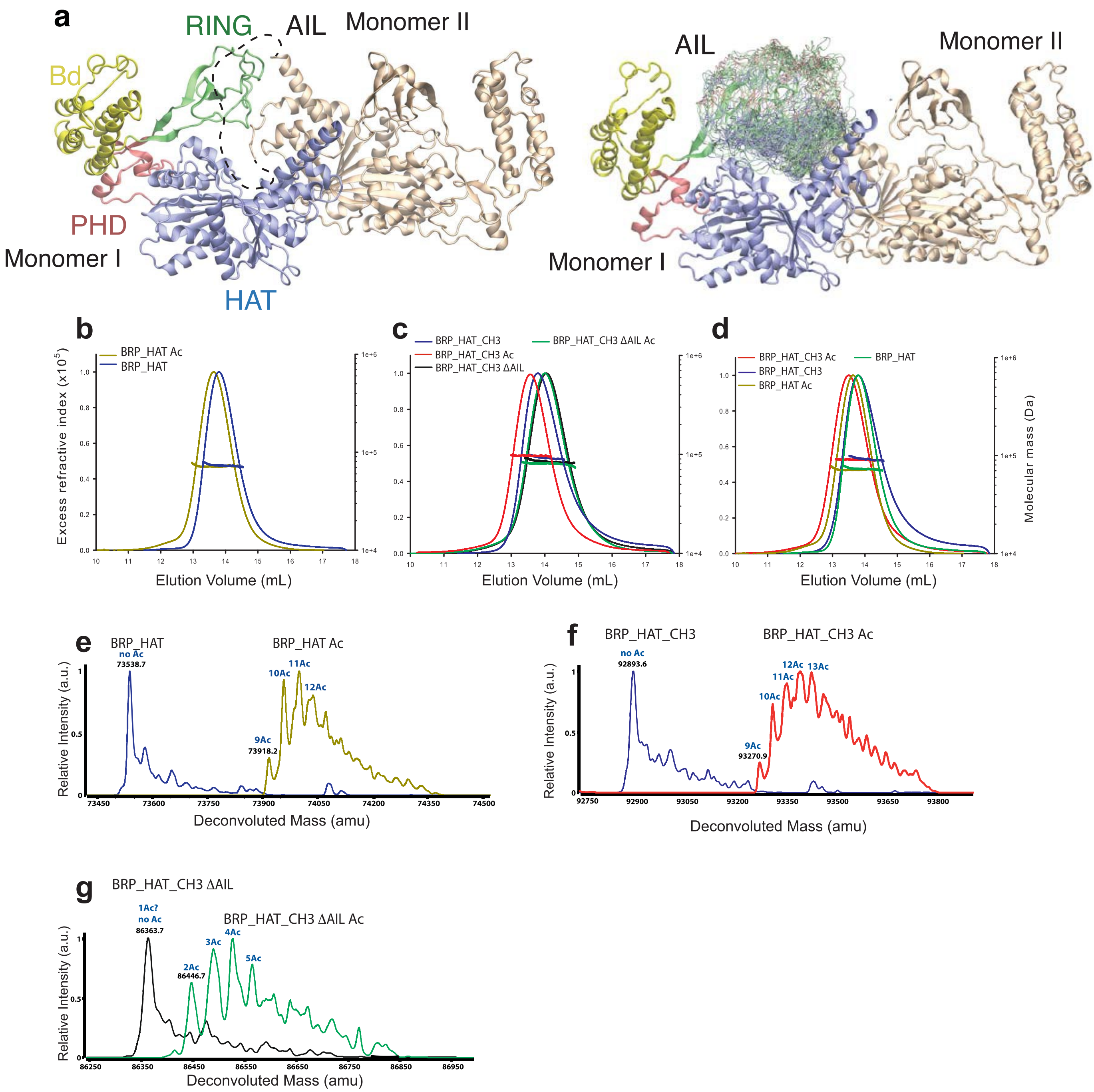
Extended Data Fig. 3 | Structural analysis of the RING domains. **a**, Superposition of the four p300 molecules (monomers I–IV) in the asymmetric crystallographic unit. Whereas the bromodomains (Bd), PHD and HAT domains superpose with a r.m.s.d. of approximately 0.9 Å, the RING domains adopt multiple conformations. **b**, $2F_o - F_c$ (blue mesh) and anomalous difference Fourier maps (orange mesh) for the four RING domains contoured around 1σ and 2.5σ, respectively.



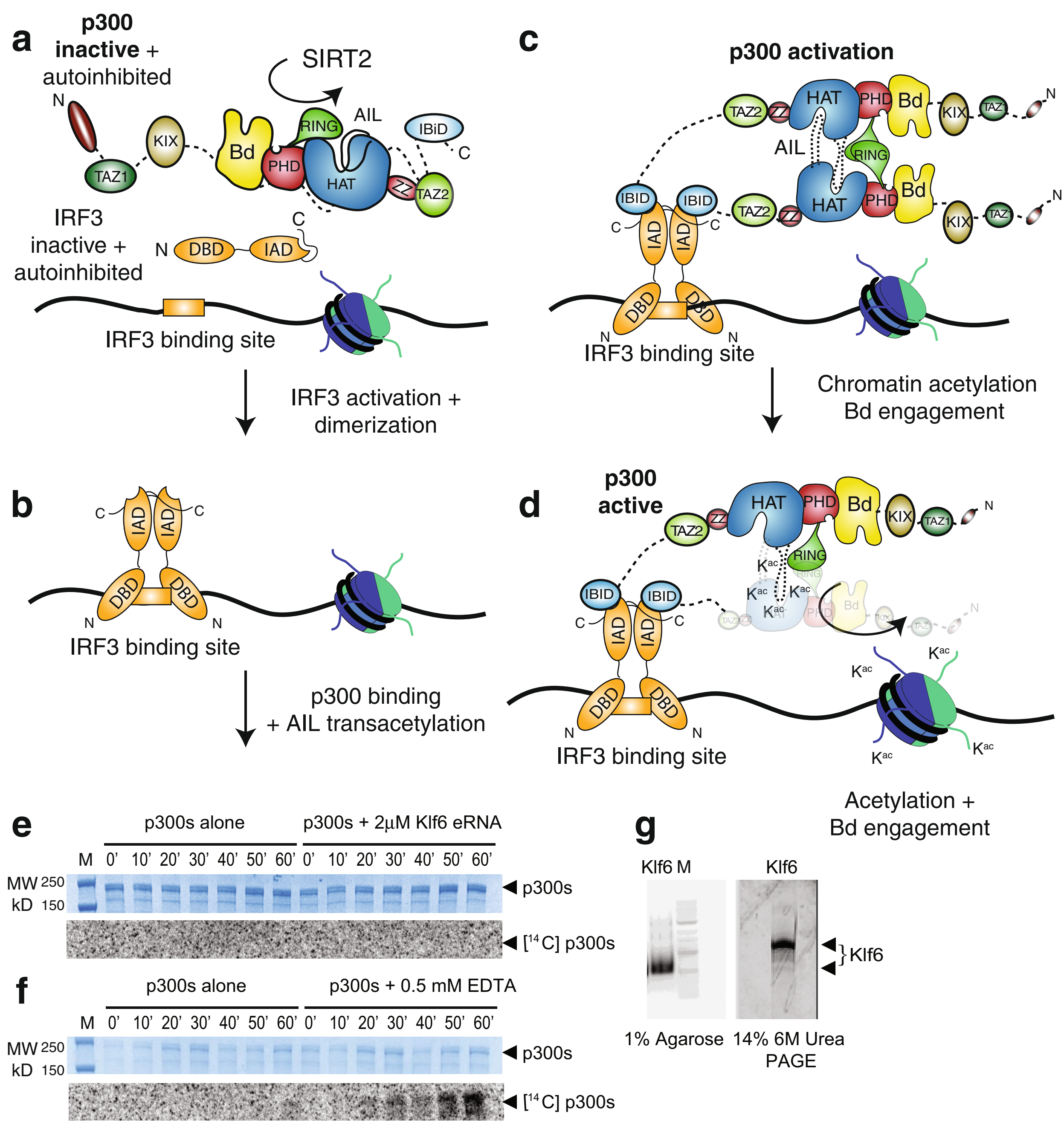
Extended Data Fig. 4 | Regulation of HAT activity by flanking domains. **a**, The domain structure of p300. Sequence conservation of the AIL is shown using WebLogo[54]. The constructs used are shown. **b**, Analysis of in vitro expression of the indicated p300 variants. Purified proteins were analysed for autoacetylation by immunoblotting with anti-p300(K1499ac) antibody (left), anti-Flag antibody (middle) and Coomassie staining (right). Representative data of three independent experiments are shown. **c**, Representative mass spectrometric analysis of BRP_HAT_ZZ_ΔAIL after in vitro expression (red) and after SIRT2 mediated deacetylation (black).



Extended Data Fig. 5 | Regulation of HAT activity by flanking domains. **a**, The AIL contributes to histone substrate acetylation of activated p300. The details of the constructs used are indicated in Extended Data Fig. 4. Defined amounts of p300 variants were incubated with acetyl-CoA and the indicated histones before SDS–PAGE analysis, followed by Coomassie staining and western blotting with the indicated antibodies. **b**, The indicated amounts of purified p300s variants were incubated with histone octamers as in **a**, followed by SDS–PAGE and immunoblot analysis with the indicated antibodies. Anti-Kac, pan-acetyl-lysine antibody. Representative data of three independent experiments are shown. **c**, Crystal structure of the H4(K12ac/K16ac) peptide bound to the BΔRP module containing an in-frame RING deletion. Amino acid residues 1169–1241 were replaced by a single glycine residue. The deletion removes the RING domain (black arrow) and does not adversely affect the structural integrity of the BΔRP module. **d**, **e**, Indicated variants of p300 were co-expressed with p53 in H1299 cells and analysed by immunofluorescence with the indicated antibodies (**d**) or by western blotting (**e**). Representative data of three independent experiments are shown. Scale bars, 10 μm.



Extended Data Fig. 6 | Autoacetylation changes the hydrodynamic properties of p300. a, Simulations of the AIL in the context of the loop-swapped dimer. Left, cartoon of the trajectory of the AIL (dashed line). Right, representative conformations with the AIL C_α backbone atoms are coloured according to charge. **b**, SEC–MALLS analysis of deacetylated (blue) and acetylated (yellow) p300 core. Note the decrease in elution volume upon acetylation. **c**, SEC-MALLS analysis of deacetylated (blue), acetylated (red) BRP_HAT_CH3 and deacetylated (black) and acetylated (green) BRP_HAT_CH3 ΔAIL. There is no increase in elution volume upon acetylation of the ΔAIL construct. **d**, Comparison of acetylated and deacetylated BRP_HAT and BRP_HAT_CH3. The deacetylated BRP_HAT (green) and deacetylated BRP_HAT_CH3 (blue) elute at the same position, which is indicative of a similar hydrodynamic radius. The acetylated BRP_HAT (yellow) and BRP_HAT_CH3 (red) elute at a larger elution volume. The normalized refractive index is plotted as a function of elution volume from an S200 column coupled to a MALLS detector. Calculated molecular masses are plotted as a function of volume for each eluted peak. The experiment was carried out at least three times with similar results. One representative example of each sample is shown. **e**, Mass spectrometry analysis (electrospray ionization) of the BRP_HAT before (blue) and after (yellow) autoacetylation. The molecular mass and the number of acetylation events are indicated. **f**, Mass spectrometry analysis of BRP_HAT_CH3 before (blue) and after (red) autoacetylation. **g**, Mass spectrometry analysis of BRP_HAT_CH3_ΔAIL before (black) and after (green) autoacetylation.



Extended Data Fig. 7 | Molecular model and controls showing that p300 acetyltransferase activity is not stimulated by eRNA. a, p300 is maintained in the inactive state by deacetylases such as SIRT2. IRF3 is autoinhibited by a C-terminal segment in the IAD domain. **b**, TBK1 phosphorylation activates and dimerizes IRF3. The activated IRF3 dimer engages the IBID domain of p300. **c**, Recruitment of two molecules of p300 results in *trans*-autoacetylation in the AIL loop and HAT activation. **d**, Activated p300 can acetylate chromatin and engage acetylated substrates via the bromodomain. **e**, A constant amount of p300s (2 μM) was incubated in [14C]acetyl-CoA alone or in the presence of 2 μM Klf6 eRNA for the indicated time points. Samples were analysed by SDS–PAGE followed by Coomassie staining (top) and autoradiography (bottom). **f**, As in **e** but in the presence of 0.5 mM EDTA. The experiment was carried out at least twice with consistency. One representative example is shown. **g**, Quality control of Klf6 RNA. 3 μg Klf6 was deposited on a 1% agarose gel or a 14% 6 M urea PAGE gel and detected by SYBR Safe stain. M, 100-bp DNA ladder.

Extended Data Table 1 | Data collection, phasing and refinement statistics

	BRP-HAT	BΔRP
Data collection		
Space group	$P2_1$	$P2_12_12_1$
Cell dimensions		
a, *b*, *c* (Å)	100.7, 146.6, 116.3	49.6, 83.7, 165.6
α, β, γ (°)	90, 91.7, 90	90, 90, 90
Resolution (Å)	50–3.10*	42.7–2.50*
No. reflections	106462	23996
R_{sym} or R_{merge}	8.7 (89.7)*	6.3 (141.3)*
$I / \sigma I$	7.48 (0.7)*	7.6 (1.0)*
Completeness (%)	99.0 (94.0)*	97.29 (92.2)*
Redundancy	1.9 (1.9)*	9.1 (3.6)*
Refinement		
Resolution (Å)	50-3.1	42.7-2.5
R_{work} / R_{free}	0.19 / 0.26	0.24 / 0.27
No. atoms	19370	2769
Protein	19100	2693
Lys-CoA ligand	256	–
Zinc	14	4
H4 K12AcK16Ac	–	76
B-factors (mean; Å^2)		
Proteins	95.0	64.0
RING domains	164.8	
Lys-CoA ligand	69.2	–
H4 K12AcK16Ac	–	60.2
R.m.s deviations		
Bond lengths (Å)	0.013	0.004
Bond angles (°)	1.73	0.68

*Data from one crystal. Values in parentheses are for highest-resolution shell.

Extended Data Table 2 | Thermodynamic analysis of the interaction between p300 BRP and histone peptides by ITC

Protein	Peptide (residues)	K_d (μM)	N*
BRP	H3 unmodified (1-20)	--	No binding
	H3 K4me1 (1-20)	--	No binding
	H3 K4me3 (1-20)	--	No binding
	H3 K9ac (1-20)	--	No binding
	H3 K14ac (1-20)	1761 ± 356	1.13
	H3 K18ac (1-20)	large	1.01
	H3 K9acK14ac (1-20)	578 ± 47	1.06
	H3 K14acK18ac (1-20)	104 ± 27	1.03
	H3 K23ac (11-30)	--	No binding
	H3 K18acK23ac (11-30)	--	No binding
	H3 S10pho (1-20)	--	No binding
	H3 T11pho (1-20)	--	No binding
BRP	H4 unmodified (4-24)	--	No binding
	H4 K5ac (4-24)	--	No binding
	H4 K8ac (4-24)	58 ± 12	1.04
	H4 K5acK8ac (4-24)	828 ± 10	1.03
	H4 K8acK12ac (4-24)	90 ± 18	1.01
	H4 K12ac (4-24)	71 ± 14	1.05
	H4 K16ac (4-24)	--	No binding
	H4 K12K16diac (4-24)	25 ± 5	0.99
	H4 K20ac (1-24)	305 ± 6	1.00
	H4 K16acK20ac (1-24)	205 ± 10	1.23
	H4 K5K8K12K16K20penta-ac (1-24)	38 ± 4	1.07
	H4 S1pho K5K8K12K16K20penta-ac (1-24)	54 ± 2	1.12
BRP_N1132A	H4 K20ac (1-24)	--	No binding
	H4 K16acK20ac (1-24)	--	No binding
	H4 K5K8K12K16K20penta-ac (1-24)	--	No binding
	H4 S1pho K5K8K12K16K20penta-ac (1-24)	--	No binding
BRP	AIL (1545-1562)	--	No binding
BRP	AIL K1549K1558K1560tri-ac (1545-1562)	--	No binding

Mean and s.d. were determined from experiments performed in triplicate. Horizontal lines separate experiments involving different histone peptides or different protein constructs.
Histone peptide sequences: H3 (1–20) ARTKQTQRKSTGGKAPRKQL, H3 (11–30) TGGKAPRKQLATKASRKSAP, H4 (4–24) GKGGKGLGKGGAKRHRKVLRD.
AIL (Autoinhibitory loop peptide) SKNAKKKNNKKTSKNKSS (1545–1562). No binding was detected to non-acetylated peptides, or with a construct containing a mutation that abolishes acetyllysine binding (N1132A).
*Binding stoichiometry.

Extended Data Table 3 | Summary of SEC-MALLS and mass spectrometry experiments

Sample	MM_{MS} Da	MM_{th} Da	Acetylation level	MM_{SLS} Da (2 mg·ml^{-1})
BRP_HAT	73538	73538	~0	73380 ± 1.5%
BRP_HAT Acetyl	73918	73538	>8	71690 ± 1.6%
BRP_HAT_CH3	92893	92891	~0	92810 ± 2.0%
BRP_HAT_CH3 Acetyl	93270	92891	>9	90700 ± 2.0%
BRP_HAT_CH3 ΔAIL	86363	86362	~0	84650 ± 2.2%
BRP_HAT_CH3 ΔAIL Acetyl	86446	86362	>2	80450 ± 1.7%

Column labelling is as follows: MM_{MS}; molar masses determined by mass spectrometry; MM_{th}, the theoretical molar mass calculated from the appropriate primary sequences, acetylation levels were estimated based on the mass differences as compared to the non-acetylated sample; MM_{SLS}, molar masses determined by SEC-MALLS at a concentration of 2 mg ml^{-1}. All p300 constructs contained the mutation Y1467F. The experiment was carried out at least three times with consistency. Results from one representative example are shown. The mass and errors reported for SEC-MALLS are the weight average molar mass and residual standard deviations of the observed data from the fitted values calculated using ASTRA.

https://doi.org/10.1038/s41586-018-0606-0

A two per cent Hubble constant measurement from standard sirens within five years

Hsin-Yu Chen[1,2]*, Maya Fishbach[2] & Daniel E. Holz[2,3,4]

Gravitational-wave detections provide a novel way to determine the Hubble constant[1–3], which is the current rate of expansion of the Universe. This 'standard siren' method, with the absolute distance calibration provided by the general theory of relativity, was used to measure the Hubble constant using the gravitational-wave detection of the binary neutron-star merger, GW170817, by the Laser Interferometer Gravitational-Wave Observatory (LIGO) and Virgo[4], combined with optical identification of the host galaxy[5,6] NGC 4993. This independent measurement is of particular interest given the discrepancy between the value of the Hubble constant determined using type Ia supernovae via the local distance ladder (73.24 ± 1.74 kilometres per second per megaparsec) and the value determined from cosmic microwave background observations (67.4 ± 0.5 kilometres per second per megaparsec): these values differ[7,8] by about 3σ. Local distance ladder observations may achieve a precision of one per cent within five years, but at present there are no indications that further observations will substantially reduce the existing discrepancies[9]. Here we show that additional gravitational-wave detections by LIGO and Virgo can be expected to constrain the Hubble constant to a precision of approximately two per cent within five years and approximately one per cent within a decade. This is because observing gravitational waves from the merger of two neutron stars, together with the identification of a host galaxy, enables a direct measurement of the Hubble constant independent of the systematics associated with other available methods. In addition to clarifying the discrepancy between existing low-redshift (local ladder) and high-redshift (cosmic microwave background) measurements, a precision measurement of the Hubble constant is of crucial value in elucidating the nature of dark energy[10,11].

We explore the expected constraints on the Hubble constant (H_0) from gravitational-wave standard sirens. The gravitational-wave data provide a direct measurement of the luminosity distance to the source, but the redshift must be determined independently. We consider gravitational-wave events both with ('counterpart') and without ('statistical') direct electromagnetic measurements of the source redshift, and carry out an end-to-end simulation of the H_0 measurement from a simulated dataset consisting of 30,000 binary neutron star (BNS) mergers and 60,000 binary black hole (BBH) mergers. We include realistic measurement uncertainties, galaxy peculiar velocities and selection effects in our analysis.

We anticipate that most, if not all, BNS mergers detected in gravitational waves will have an electromagnetic counterpart (for example, from associated isotropic[12,13] kilonova emission[5,6]) that will allow a unique host galaxy identification[4]. Assuming that the BNS population is similar to the population of short γ-ray bursts, we expect the typical offset between a kilonova and its associated host galaxy to be no more than[14] 100 kpc. Since Advanced LIGO–Virgo BNS detections will be within 400 Mpc, it will be possible to identify host galaxies as faint as $0.003L_B^*$ (apparent magnitudes <23), where L_B^* is the B-band Schechter function parameter (see Methods), with modest observational resources. We find that in this counterpart case, the fractional H_0 uncertainty will scale roughly as $15\%/\sqrt{N}$, where N is the number of BNS mergers detected by the LIGO-Hanford, LIGO-Livingston and Virgo network (HLV). Throughout, we quote fractional H_0 measurement uncertainties defined as half the width of the symmetric 68% credible interval divided by the median. If the Kamioka Gravitational Wave Detector (KAGRA) and LIGO-India join the detector network (HLVJI), this convergence improves slightly to $13\%/\sqrt{N}$, because a five-detector network tends to provide better measurements of the source inclination, and therefore distance, owing to the improved polarization information.

We note that the representative fractional H_0 measurement uncertainty σ_{H_0} (15% for the three-detector network, and 13% for the five-detector network), is smaller than the typical width of the H_0 measurement from an individual event (GW170817 provided an unusually tight measurement; see Extended Data Fig. 1). This is because for a single event, the H_0 posterior probability density function is a highly non-Gaussian function; the distance–inclination degeneracy leads to long tails up to large distances (and low H_0 values) for edge-on sources, and tails in the opposite direction for face-on sources. Combining these asymmetric distributions leads to a $1/\sqrt{N}$ convergence with a smaller effective σ_{H_0} than the width of a typical single-event H_0 measurement[15,16]. Furthermore, owing to the asymmetry of the single-event measurements, it may take about 20 events to reach the expected $1/\sqrt{N}$ convergence rate. For example, we may get lucky in the first few events and get an unusually good H_0 measurement (GW170817 is an excellent example of this), after which we will converge more slowly than $1/\sqrt{N}$ for some time as we detect average events. After about 20 events, however, we will have a sufficient statistical sample of detections to have converged to a representative σ_{H_0} for the population. At this point, the combined H_0 measurement approaches a Gaussian distribution and we reach the expected $1/\sqrt{N}$ behaviour.

To predict how the H_0 measurement improves with time, we consider the BNS rates inferred[17] from GW170817, $1{,}540^{+3{,}200}_{-1{,}220}$ Gpc^{-3} yr^{-1}, together with the planned network sensitivity and duty cycle, to compute the expected number of detections at each observing stage. Figure 1 shows the improvement in the H_0 measurement as BNS detections with unique host galaxies (and associated redshifts) are accumulated. We start with a 15% prior measurement on H_0, representing the constraint[4] from GW170817; we approximate this by a Gaussian centred at 67.8 km s^{-1} Mpc^{-1} with a standard deviation of 10.2 km s^{-1} Mpc^{-1}, but the exact centre and shape of the H_0 posterior do not affect our results. The rate of detections will increase as the gravitational-wave network improves in sensitivity between LIGO-Virgo's third observing run ('O3 HLV'), HLV at design sensitivity ('Design HLV'), and the five-detector network ('Design HLVJI'), from an average of five BNS detections per year in O3 HLV, to 32 and 39 detections per year for Design HLV and Design HLVJI, respectively. The merger rate provides the major source of uncertainty in predicting the H_0 measurement error. The solid line in Fig. 1b shows the average H_0 measurement error over 100 realizations assuming the median BNS merger rate, while the lower/upper bounds of the shaded band assume the upper/lower 90% bounds, respectively, on the merger rate inferred from GW170817.

[1]Black Hole Initiative, Harvard University, Cambridge, MA, USA. [2]Department of Astronomy and Astrophysics, University of Chicago, Chicago, IL, USA. [3]Enrico Fermi Institute, Department of Physics and Kavli Institute for Cosmological Physics, University of Chicago, Chicago, IL, USA. [4]Physics Department and Kavli Institute for Particle Astrophysics and Cosmology, Stanford University, Stanford, CA, USA. *e-mail: hsinyuchen@fas.harvard.edu

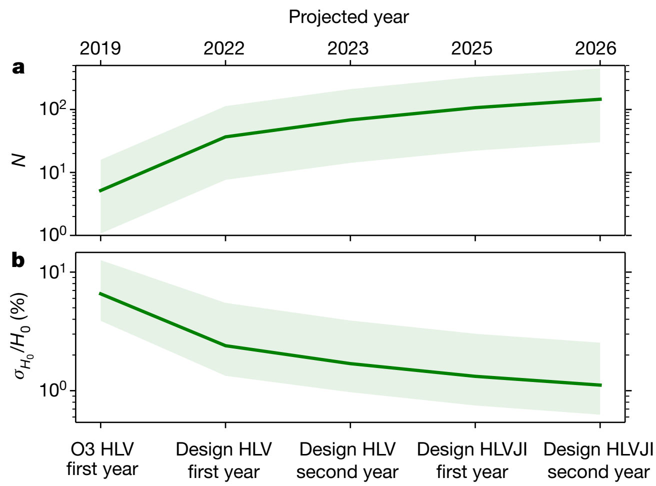


Fig. 1 | Projected number of BNS detections and corresponding fractional error for the standard siren H_0 measurement. a, The expected total number N of BNS detections for future observing runs, using the median merger rate (solid green curve) and upper and lower rate bounds (shaded band). **b**, The corresponding H_0 measurement error, defined as half of the width of the 68% symmetric credible interval divided by the posterior median. The band corresponds to the uncertainty in the merger rate shown in **a**. These measurements assume an optical counterpart, and the associated redshift, for all BNS systems detected with gravitational waves.

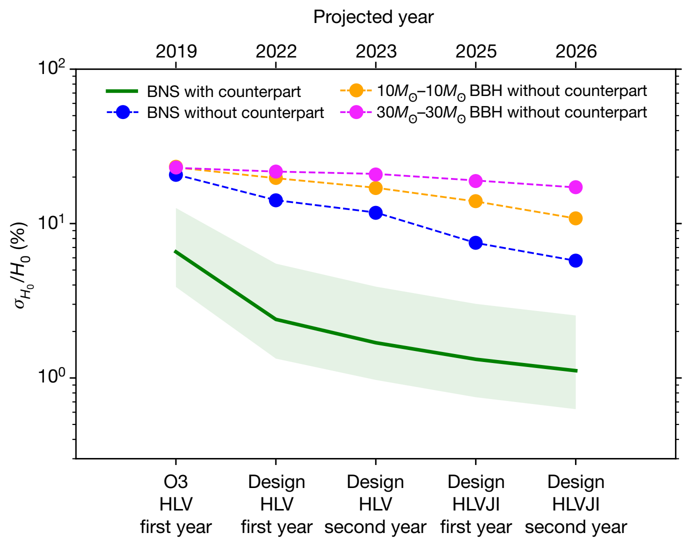


Fig. 2 | Projected fractional error for the standard siren H_0 measurement for BNSs and BBHs for future gravitational-wave detector networks. The green shaded band corresponds to the BNS rate uncertainty (see Fig. 1); the same uncertainty applies to the 'BNS without counterpart' curve. For the 'without counterpart' curves, we adopt a statistical standard siren approach using only events localized to within 10,000 Mpc^3; events with larger volumes do not contribute noticeably (see the text). Constraints from BBH systems without counterparts are inferior, despite higher rates, owing to the larger numbers of potential host galaxies per event.

We find that, if it is possible to independently measure a unique redshift for all BNS events, the fractional uncertainty on H_0 will reach 2% (at the 1σ level) by the end of two years of HLV at design sensitivity (in about 2023; corresponding to about 50 events), sufficient to arbitrate the current tension between local and high-z measurements of H_0. After about 100 BNS events, gravitational-wave standard sirens would provide a 1% determination of H_0. This is expected to happen after about two years of operation of the full HLVJI network (around 2026), but given the rate uncertainties, it could happen many years later, or could happen as early as 2023.

Not all sources will have associated transient electromagnetic counterparts: we may fail to identify the counterparts to some BNS mergers, and counterparts are not expected for BBH mergers. For cases where a unique counterpart cannot be identified, it is possible to carry out a measurement of the Hubble constant using the statistical approach. To do this, the redshifts of all potential host galaxies within the gravitational-wave three-dimensional localization region are incorporated, yielding an H_0 measurement that is inferior to what can be calculated using a counterpart, but is still informative once many detections are combined. This means that, in the absence of a counterpart, only those gravitational-wave events with small enough localization volumes yield informative H_0 measurements. If the localization volume is too large, it contains a large number of potential host galaxies, which will largely wash out the contribution from the correct host galaxy. Additionally, it may be difficult in practice to construct a complete galaxy catalogue over a large volume with precise galaxy redshifts. We find that for BNSs without counterparts, combining the H_0 measurement from events that are localized to within 10,000 Mpc^3 (approximately 40% of events) yields identical constraints to the combined measurement using the full sample—events localized to greater than 10,000 Mpc^3 do not contribute to the measurement. For this reason, we use only the sources localized to within 10,000 Mpc^3 for the no-counterpart projections in Fig. 2. We note that for all of the no-counterpart curves in Fig. 2, we start with a flat H_0 prior in the range 50–100 km s^{-1} Mpc^{-1}.

Because BBH systems tend to have much larger localization volumes than BNS systems (as they are more massive and found at greater distances), the statistical H_0 measurement for BBHs converges very slowly, even though they are detected at higher rates. We consider both 'light' (components of mass $10M_{\odot}$, where $M_{\odot}$ is the solar mass; denoted as '$10M_{\odot}$–$10M_{\odot}$') and 'heavy' (components of mass $30M_{\odot}$; denoted as '$30M_{\odot}$–$30M_{\odot}$') BBHs, assuming merger rates of 80^{+90}_{-70} Gpc^{-3} yr^{-1} for the $10M_{\odot}$–$10M_{\odot}$ BBHs and 11^{+13}_{-10} Gpc^{-3} yr^{-1} for the $30M_{\odot}$–$30M_{\odot}$ BBHs[18] (see Methods). Only about 3% of the light BBHs and about 0.5% of the heavy BBHs are localized to within 10,000 Mpc^3, which means that we expect to detect only 16^{+19}_{-14} well localized BBHs by 2026. This leads to an approximately 10% H_0 measurement with BBHs by 2026. We note that the constraints from statistical BBH standard sirens improve if the BBH rates are on the high end, as well as if the BBH mass function favours low masses.

For the projections in Fig. 2, we assumed that galaxies are distributed uniformly in a comoving volume and that complete catalogues are available. If we incorporate the clustering of galaxies due to large-scale structure, the convergence rate in the statistical case improves by a factor of about 2.5 (see Methods). Incorporating this large-scale structure effect, we find that we will still need to detect more than about 50 BNSs without a counterpart to reach a 6% H_0 measurement, compared to only ten BNSs or fewer with a counterpart. Meanwhile, accounting for galaxy catalogue incompleteness provides an additional source of uncertainty (see equation (5) in Methods), which can cancel out some of the improvement due to large-scale structure. For example, for a galaxy catalogue completeness of 50%, the H_0 measurement would be degraded by about a factor of two. Therefore, incorporating the effects of large-scale structure and catalogue incompleteness, we expect that in practice the H_0 constraints in the statistical case will be slightly better than our prediction in Fig. 2, where the precise factor depends on properties of the relevant host galaxies and completeness of the catalogue.

Besides the with-counterpart and without-counterpart cases, we can also anticipate a situation in which we have a counterpart detection but no unambiguous host association. For example, an optical counterpart could be relatively isolated on the sky without a clearly identified host galaxy, or may have multiple possible host galaxies. In this case we can pursue a pencil-beam strategy, for example, focusing on the volume within 100 kpc of the counterpart (see Methods). For BNSs this will reduce the relevant volume to about 10 Mpc^3, for which we expect to have only about one potential host galaxy or galaxy group, which thereby reduces to the with-counterpart case.

In addition to the BNSs and BBHs discussed here, we can expect mergers of a neutron star and a black hole[19–21], and these may have detectable

electromagnetic counterparts. Although the rates for these systems are uncertain and expected to be low, they will also be seen to greater distances than BNS systems, which may render them useful as standard sirens[22].

We note that our measurements of distance do not use any astrophysical modelling. However, associated electromagnetic observations (for example, from short γ-ray burst afterglows or jet breaks) can provide additional constraints on the inclination, and thereby improve the individual measurements[4,23] of distance, leading to a tighter measurement of H_0. In this sense our counterpart results can be considered a conservative estimate. However, one of the advantages of standard sirens is that they are 'pure' measurements of luminosity distance, avoiding complicated astrophysical distance ladders or poorly understood calibration processes, and instead are calibrated directly by the theory of general relativity to cosmological distances. By introducing additional constraints based on astronomical observations (for example, independent beaming measurements or estimates of the mass distribution or equation of state of neutron stars), there is the potential to introduce systematic biases that could fundamentally contaminate the standard siren measurements. In the present analysis we do not consider these additional constraints, although they may indeed have an important part to play in future standard siren science.

Eventually systematic errors in the amplitude calibration of the detectors may become a source of concern, because the luminosity distance is encoded in the amplitude of the gravitational-wave signal. However, the calibration uncertainty is currently limited by the photon calibrator to around 1%, and this is likely to improve[24]; we look forward to an era where sub-1% calibration becomes a necessity, but this is a number of years away. Another possible source of distance uncertainty is gravitational lensing. However, at the typical redshift of BNS and BBH systems ($z < 0.5$ at design sensitivity) the effect will be minor relative to the uncertainty from the distance–inclination degeneracy[25]. In addition, for sufficient numbers of sources the effects of lensing will average away[15]. Of course, gravitational-wave cosmology is a new field, and unforeseen systematics could certainly arise as we push our measurements to the 1% level and beyond.

We stress that our projected H_0 constraints are subject to several important uncertainties, the largest one of which is the merger rate of BNS and BBH systems. The detection rate for BBHs depends sensitively on the mass distribution, which is not currently well constrained[26,27]. Future detections will bring a better understanding of the merger rates and mass distributions of compact objects, allowing for improved predictions. Nevertheless, it is clear that gravitational-wave standard sirens will provide precision constraints on cosmology in the upcoming advanced-detector era of gravitational-wave astronomy.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0606-0.

Received: 9 January; Accepted: 16 August 2018;
Published online 17 October 2018.

1. Schutz, B. F. Determining the Hubble constant from gravitational wave observations. *Nature* **323**, 310–311 (1986).
2. Holz, D. E. & Hughes, S. A. Using gravitational-wave standard sirens. *Astrophys. J.* **629**, 15 (2005).
3. Dalal, N., Holz, D. E., Hughes, S. A. & Jain, B. Short GRB and binary black hole standard sirens as a probe of dark energy. *Phys. Rev. D* **74**, 063006 (2006).
4. Abbott, B. P. et al. A gravitational-wave standard siren measurement of the Hubble constant. *Nature* **551**, 85–88 (2017).
5. Coulter, D. A. et al. Swope Supernova Survey 2017a (SSS17a), the optical counterpart to a gravitational wave source. *Science* **358**, 1556–1558 (2017).
6. Soares-Santos, M. et al. The electromagnetic counterpart of the binary neutron star merger LIGO/Virgo GW170817. I. Discovery of the optical counterpart using the dark energy camera. *Astrophys. J.* **848**, L16 (2017).
7. Riess, A. G. et al. A 2.4% determination of the local value of the Hubble constant. *Astrophys. J.* **826**, 56 (2016).
8. Aghanim, N. et al. Planck 2018 results. VI. Cosmological parameters. Preprint at https://arxiv.org/abs/1807.06209 (2018).
9. Riess, A. G. et al. New parallaxes of galactic Cepheids from spatially scanning the Hubble Space Telescope: implications for the Hubble constant. *Astrophys. J.* **855**, 136 (2018).
10. Hu, W. Dark energy probes in light of the CMB. *ASP Conf. Series* **339**, 215 (2005); preprint at https://arxiv.org/abs/astro-ph/0407158.
11. Di Valentino, E., Holz, D. E., Melchiorri, A. & Renzi, F. The cosmological impact of future constraints on H_0 from gravitational-wave standard sirens. Preprint at https://arxiv.org/abs/1806.07463 (2018).
12. Li, L.-X. & Paczyński, B. Transient events from neutron star mergers. *Astrophys. J.* **507**, L59–L62 (1998).
13. Metzger, B. D. et al. Electromagnetic counterparts of compact object mergers powered by the radioactive decay of r-process nuclei. *Mon. Not. R. Astron. Soc.* **406**, 2650–2662 (2010).
14. Fong, W. & Berger, E. The locations of short gamma-ray bursts as evidence for compact object binary progenitors. *Astrophys. J.* **776**, 18 (2013).
15. Hirata, C. M., Holz, D. E. & Cutler, C. Reducing the weak lensing noise for the gravitational wave Hubble diagram using the non-Gaussianity of the magnification distribution. *Phys. Rev. D* **81**, 124046 (2010).
16. Nissanke, S. et al. Determining the Hubble constant from gravitational wave observations of merging compact binaries. Preprint at https://arxiv.org/abs/1307.2638 (2013).
17. Abbott, B. P. et al. GW170817: observation of gravitational waves from a binary neutron star inspiral. *Phys. Rev. Lett.* **119**, 161101 (2017).
18. Abbott, B. P. et al. GW170104: observation of a 50-solar-mass binary black hole coalescence at redshift 0.2. *Phys. Rev. Lett.* **118**, 221101 (2017).
19. Dominik, M. et al. Double compact objects III: gravitational-wave detection rates. *Astrophys. J.* **806**, 263 (2015).
20. Belczynski, K. et al. Compact binary merger rates: comparison with LIGO/Virgo upper limits. *Astrophys. J.* **819**, 108 (2016).
21. Belczynski, K., Holz, D. E., Bulik, T. & O'Shaughnessy, R. The first gravitational-wave source from the isolated evolution of two stars in the 40–100 solar mass range. *Nature* **534**, 512–515 (2016).
22. Vitale, S. & Chen, H.-Y. Measuring the Hubble constant with neutron star black hole mergers. *Phys. Rev. Lett.* **121**, 021303 (2018).
23. Guidorzi, C. et al. Improved constraints on H_0 from a combined analysis of gravitational-wave and electromagnetic emission from GW170817. *Astrophys. J.* **851**, L36 (2017).
24. Karki, S. et al. The Advanced LIGO photon calibrators. *Rev. Sci. Instrum.* **87**, 114503 (2016).
25. Holz, D. E. & Linder, E. V. Safety in numbers: gravitational lensing degradation of the luminosity distance-redshift relation. *Astrophys. J.* **631**, 678–688 (2005).
26. Abbott, B. P. et al. The rate of binary black hole mergers inferred from advanced LIGO observations surrounding GW150914. *Astrophys. J.* **833**, L1 (2016).
27. Fishbach, M. & Holz, D. E. Where are LIGO's big black holes? *Astrophys. J.* **851**, L25 (2017).

Acknowledgements We acknowledge discussions with L. Blackburn, R. Essick, W. Farr and J. Gair. We were supported in part by NSF CAREER grant PHY-1151836 and NSF grant PHY-1708081. We were also supported by the Kavli Institute for Cosmological Physics at the University of Chicago through NSF grant PHY-1125897 and an endowment from the Kavli Foundation. We acknowledge the University of Chicago Research Computing Center for support of this work. H.-Y.C. was supported in part by the Black Hole Initiative at Harvard University, through a grant from the John Templeton Foundation. M.F. was supported by the NSF Graduate Research Fellowship Program under grant DGE-1746045.

Author contributions H.-Y.C. led the project, conducted the simulations and led the analysis. M.F. provided the mathematical derivations and contributed to the analysis and results. D.E.H. conceived the project, supervised the research, and contributed to the analysis and results. All authors contributed to the draft preparation.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0606-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to H.-Y.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

We present our method for inferring cosmological parameters from gravitational-wave (GW) and electromagnetic (EM) measurements. We first simulate a representative sample of GW detections for a range of detector configurations. We then simulate the analysis of these data sets, and explore the resulting standard siren constraints. In what follows we highlight important aspects of our calculation, such as the role of peculiar velocities and selection effects.

Synthetic events and host galaxies. Measuring H_0 with standard sirens relies on our ability to extract the luminosity distance and sky position of GW sources. We follow the procedure in ref. [28] to localize synthetic BNS merger and BBH merger detections. The population of binaries are distributed uniformly in comoving volume in a Planck (2015)[46] cosmology ($\Omega_{M_0} = 0.308$, $\Omega_{\Lambda_0} = 0.692$, $H_0 = 67.8$). We assume that the BNS merger rate follows the rate measured in ref. [17]. To estimate the merger rate of $10M_\odot$–$10M_\odot$ and $30M_\odot$–$30M_\odot$ BBHs from the rate measured in ref. [18], we assume that the BBH mass function follows a Salpeter power law and use $10M_\odot$–$10M_\odot$ BBHs to characterize all BBHs with primary component masses between $5M_\odot$ and $15M_\odot$, and $30M_\odot$–$30M_\odot$ BBHs to characterize all BBHs with primary component masses between $20M_\odot$ and $50M_\odot$. We do not place additional cuts on the secondary masses, which are distributed uniformly between $5M_\odot$ and the primary mass.

The detection rate of sources depends on the sensitivity, observing time and duty cycle of the GW detector network. We assume that the LIGO–Virgo network operates for one year at projected O3 sensitivity, followed by two one-year-long observing runs of LIGO-Hanford+ LIGO-Livingston+Virgo (HLV) at design sensitivity and two one-year-long runs of the five-detector network, LIGO-Hanford+ LIGO-Livingston+Virgo+KAGRA+LIGO-India (HLVJI), at design sensitivity[29]. We take the combined duty cycle to be 0.5 for the HLV detector configuration and 0.3 for HLVJI. The number of detections is subject to Poisson statistics, and we simulate detections according to the merger rate, network sensitivity, observing time and duty cycle.

To determine whether a binary merger is detected, we calculate the matched-filter signal-to-noise ratio (SNR) for each simulated binary. We draw the 'measured' SNR from a Gaussian distribution centred at the matched-filter value with a standard deviation of $\sigma = 1$. Binary mergers are detected only if their measured network SNR is greater than 12. For each detected merger, we calculate its three-dimensional localization according to the methods in ref. [28] (We have verified that this procedure yields results which are consistent with the full parameter estimation pipeline, LALInference[30].) The three-dimensional localization takes the form of a posterior probability distribution function, $p(\alpha, \delta, D_L|d_{\mathrm{GW}})$, over the sky position (α, δ) and luminosity distance, D_L, given the GW data, d_{GW}.

The GW signal from each detected binary merger provides a measurement of D_L. To calculate H_0, we must also measure a redshift for each binary merger. Throughout, we take the redshift, z, to be the peculiar-velocity-corrected redshift; that is, the redshift that the source would have if it were in the Hubble flow. We consider two cases: the redshift information either comes from a direct EM counterpart, such as a short γ-ray burst/afterglow and/or a kilonova ('with-counterpart'), or a statistical analysis over a catalogue of potential host galaxies ('statistical').

In the with-counterpart case, we assume that the EM counterpart is close enough to its host galaxy that the host can be unambiguously identified, and we can measure its sky position and redshift. This is a reasonable assumption based on the distribution of offsets between short γ-ray bursts and their host galaxies, assuming that short γ-ray bursts trace a population similar to that of BNS mergers, and taking into account that detected BNS mergers will be at much lower redshifts than the short γ-ray burst population. We assume that the sky position of each host galaxy is perfectly measured (that is, with negligible measurement error), meaning we can fix the source sky position to the location of the counterpart in the GW parameter estimation (rather than marginalizing over all sky positions). The GW distance posterior changes slowly over the sky and therefore is not sensitive to the precise location of the counterpart. However, since the GW sky localization areas can be very large, fixing the source position can lead to important improvements in the distance, and hence H_0, measurements. We also assume that the peculiar-velocity-corrected redshift, z, is measured with a 1σ error of 200 km s^{-1} c^{-1}, where c is the speed of light in vacuum, which is a typical uncertainty for the peculiar velocity correction[31,32].

In the absence of an EM counterpart we cannot identify a single host galaxy, and must use a catalogue of all potential host galaxies[1,33]. To simulate the galaxy catalogues we consider two cases: a uniform-in-comoving-volume distribution of galaxies, and a distribution that follows the large-scale structure as simulated by the MICE galaxy catalogue[34–37]. In the uniform distribution case, we construct a mock catalogue by distributing galaxies uniformly in comoving volume with a number density of 0.02 Mpc^{-3}. This corresponds to the number density of galaxies that are 25% as bright as the Milky Way, assuming the galaxy luminosity function is described by the Schechter function[38] with B-band parameters $\phi^*_{\mathrm{B}} = 1.6\times10^{-2}h^3$ Mpc^{-3}, $\alpha_{\mathrm{B}} = -1.07$, $L^*_{\mathrm{B}} = 1.2\times10^{10}h^{-2}L^\odot_{\mathrm{B}}$, and $h = 0.7$ (where $L^\odot_{\mathrm{B}}$ is the solar luminosity in the B band), and integrating down to $0.16L^*_{\mathrm{B}}$ to find the luminosity density. (This corresponds to 83% of the total luminosity[39–41].) The lower luminosity limit of the MICE catalogue is similar. Thus, we assume that only galaxies brighter than $0.16L^*_{\mathrm{B}}$ can host binary mergers, although we note that the population of host galaxies is currently uncertain, and we can modify the assumed luminosity limit by including the effects of catalogue incompleteness. A lower luminosity limit would increase the galaxy density and weaken the H_0 constraints in the statistical case.

H_0 uncertainty. Not all GW events contribute equally to the H_0 measurements. In the counterpart case, the fractional error on the H_0 measurement from a single source depends on the fractional distance uncertainty of the GW source and the fractional redshift uncertainty of its host galaxy. To first order, this is:

$$\left(\frac{\sigma_{H_0}}{H_0}\right)^2\Bigg|_{\mathrm{1gal}} \approx \left(\frac{\sigma_{v_H}}{v_H}\right)^2 + \left(\frac{\sigma_{D_L}}{D_L}\right)^2 \quad (1)$$

where '1gal' denotes the case of a uniquely identified host galaxy and v_H is the peculiar-velocity-corrected 'Hubble velocity'. Because the recessional velocity uncertainty, σ_{v_H}, is typically around 150–250 km s^{-1}, the fractional recessional velocity decreases with distance. Meanwhile, the fractional distance uncertainty scales roughly inversely with SNR, and therefore tends to increase with distance. There is thus a 'sweet spot', at which the peculiar velocities and the distance uncertainties are comparable; for LIGO–Virgo's second observing run; this was about 30 Mpc, near the distance of GW170817. The distance of the sweet spot will increase as the networks become more sensitive; for detectors at design sensitivity the ideal BNS standard siren distance will be about 50 Mpc. At distances beyond this, the distance uncertainty will tend to dominate the peculiar velocity uncertainty; in this regime, the nearest (highest SNR) events tend to provide the tightest H_0 constraints. This can be seen in Extended Data Fig. 1, which shows the fractional H_0 uncertainty for individual events, plotted against the median posterior distance and 90% posterior localization volumes. However, we note that the relationship between median distance, localization volumes, and fractional H_0 uncertainty is not very tight. Prior to identifying the counterpart for a particular event, we can estimate the accuracy of the H_0 measurement from the width and central value (for example, median) of the GW distance posterior according to equation (1), using an estimated $v_H \approx 70\langle D_L\rangle$ km s^{-1} Mpc^{-1}, where $\langle D_L\rangle$ is the median GW distance. (Here we must use the GW posterior marginalized over the sky position, as we do not yet know the sky position of the counterpart.) We verify that this estimate of the combined distance and redshift uncertainty is a reasonable proxy for the resulting H_0 uncertainty, assuming that an EM counterpart is found and provides an independent measurement of redshift.

In the absence of a counterpart, we cannot assign a unique host, and so the H_0 error increases with the number of potential host galaxies in the localization volume. Galaxy clustering can mitigate this, as we discuss in the main text. For example, in the case of GW170817, the optical counterpart was found in NGC 4993, which is a member of a group of about 20 galaxies, all of which have an equivalent Hubble recessional velocity[42]. On the other hand, catalogue incompleteness degrades the H_0 measurement, as we have to consider an additional background of uniformly distributed galaxies (see equation (5)).

Bayesian model. For a single event with GW and EM data, d_{GW} and d_{EM}, we can write the likelihood as:

$$p(d_{\mathrm{GW}}, d_{\mathrm{EM}}|H_0) = \frac{\int p(d_{\mathrm{GW}}, d_{\mathrm{EM}}, D_L, \alpha, \delta, z|H_0)\mathrm{d}D_L\mathrm{d}\alpha\mathrm{d}\delta\mathrm{d}z}{\beta(H_0)} \quad (2)$$

where we have included a normalization term in the denominator, $\beta(H_0)$, to account for selection effects and ensure that the likelihood integrates to unity. We can factor the numerator in equation (2) as:

$$\int p(d_{\mathrm{GW}}, d_{\mathrm{EM}}, D_L, \alpha, \delta, z|H_0)\mathrm{d}D_L\mathrm{d}\alpha\mathrm{d}\delta\mathrm{d}z$$

$$= \int p(d_{\mathrm{GW}}|D_L, \alpha, \delta)p(d_{\mathrm{EM}}|z, \alpha, \delta)p(D_L|z, H_0)p_0(z, \alpha, \delta|H_0)\mathrm{d}D_L\mathrm{d}\alpha\mathrm{d}\delta\mathrm{d}z$$

$$= \int p(d_{\mathrm{GW}}|D_L, \alpha, \delta)p(d_{\mathrm{EM}}|z, \alpha, \delta)\delta(D_L - \hat{D}_L(z, H_0))p_0(z, \alpha, \delta|H_0)\mathrm{d}D_L\mathrm{d}\alpha\mathrm{d}\delta\mathrm{d}z$$

$$= \int p(d_{\mathrm{GW}}|\hat{D}_L(z, H_0), \alpha, \delta)p(d_{\mathrm{EM}}|z, \alpha, \delta)p_0(z, \alpha, \delta|H_0)\mathrm{d}\alpha\mathrm{d}\delta\mathrm{d}z \quad (3)$$

where $\hat{D}_L(z, H_0)$ denotes the luminosity distance of a source at redshift z, given a Hubble constant of H_0 and leaving all other cosmological parameters fixed to the Planck values[46] ($\Omega_{M_0} = 0.308$, $\Omega_{\Lambda_0} = 0.692$). We can, alternatively, marginalize

over these other cosmological parameters, but since most detected binaries will be at low redshifts, the effects of other cosmological parameters on the z–D_L relation are small. The term $p(d_{GW}|D_L,\alpha,\delta)$ is the marginalized likelihood of the GW data given a compact binary source at distance D_L and sky position (α, δ), marginalized over all other parameters. Throughout, we assume that we can construct a catalogue of the potential host galaxies for each event, and take the prior $p_0(z,\alpha,\delta|H_0)$ to be a sum of Gaussian distributions centred at the measured redshifts and sky positions of the galaxies:

$$p_0(z,\alpha,\delta|H_0) = p_{\text{catalogue}}(z,\alpha,\delta) = \frac{1}{N_{\text{gal}}}\sum_i^{N_{\text{gal}}} N[\bar{z}^i,\sigma_z^i](z)N[\bar{\alpha}^i,\sigma_\alpha^i](\alpha)N[\bar{\delta}^i,\sigma_\delta^i](\delta) \tag{4}$$

In practice, we ignore the uncertainties on the measured sky coordinates and treat the Gaussian distributions as δ-functions centred at the measured $\bar{\alpha}^i$, $\bar{\delta}^i$ We take $\bar{z}^i$ to be the peculiar-velocity-corrected redshifts, and assume a standard deviation of $c\sigma_z^i = 200$ km s^{-1} for each. In the above, we assign equal weights to each galaxy in the catalogue, but if we believe that certain galaxies are a priori more likely to be GW hosts, we can assign weights accordingly. For example, we can weight the galaxies in equation (4) by their stellar or star-forming luminosity, or by some assumed redshift-dependent coalescence rate of the GW sources. A critical assumption is that the sum in equation (4) contains the correct host galaxy. If we believe the catalogue is incomplete, we must replace our prior, $p_0(z)$, with a weighted sum containing both the observed galaxies, equation (4)—weighted by the overall completeness fraction of the catalogue—and a smooth, uniform-in-comoving-volume distribution accounting for the unobserved galaxies:

$$p_0(z,\alpha,\delta|H_0) = fp_{\text{catalogue}}(z,\alpha,\delta) + (1-f)p_{\text{miss}}(z,\alpha,\delta|H_0) \tag{5}$$

where $p_{\text{catalogue}}$ is given by equation (4), and:

$$p_{\text{miss}}(z,\alpha,\delta|H_0) \propto [1-P_{\text{complete}}(z,\alpha,\delta)]\frac{\mathrm{d}V_c}{\mathrm{d}z\mathrm{d}\alpha\mathrm{d}\delta} \tag{6}$$

where $P_{\text{complete}}(z,\alpha,\delta)$ is the probability of a galaxy at (z,α,δ) being in the catalogue. Meanwhile the completeness fraction f is given by:

$$f = \frac{1}{V_c(z_{\max})}\int_\alpha\int_\delta\int_0^{z_{\max}} P_{\text{complete}}(z,\alpha,\delta)\frac{\mathrm{d}V_c}{\mathrm{d}z\mathrm{d}\alpha\mathrm{d}\delta}\mathrm{d}z\mathrm{d}\alpha\mathrm{d}\delta \tag{7}$$

where $z_{\max}$ is the maximum galaxy redshift considered in the analysis of an individual event, and $V_c(z_{\max})$ is the total comoving volume enclosed within $z_{\max}$.

In the case where we have an EM counterpart, the likelihood $p(d_{EM}|z,\alpha,\delta)$ picks out one of the galaxies in the catalogue, so that the sum in the prior reduces to a single term corresponding to the EM-identified host galaxy. In the case where there is no EM counterpart, the EM data are uninformative, and we set the likelihood $p(d_{EM}|z,\alpha,\delta) \propto$ constant. In the case where we have an EM counterpart but cannot pick out a unique host galaxy, one could consider a 'pencil beam' containing all the potential host galaxies within around 100 kpc in projected distance on the sky. We assume the sky position of the counterpart is perfectly measured to be $(\bar{\alpha},\bar{\delta})$, and take the term $p(d_{EM}|z,\alpha,\delta)$ to be a top hat that picks out all of the galaxies within some angular radius r (corresponding to around 100 kpc in projected distance) of the counterpart's sky position. Thus, the numerator of equation (3) reduces to:

$$\int_{\langle(\alpha,\delta)|(\bar{\alpha},\bar{\delta})\rangle<r} p(d_{GW}|\hat{D}_L(z,H_0),\bar{\alpha},\bar{\delta})p_0(z,\alpha,\delta)\mathrm{d}z\mathrm{d}\alpha\mathrm{d}\delta \tag{8}$$

and we sum over all galaxies within about 100 kpc in projected distance, but no longer weight them by the likelihood of the GW source at the corresponding sky position. Alternatively, we can incorporate assumptions about the kick distribution in the form of $p(d_{EM}|z,\alpha,\delta)$ and place more weight at galaxies close to $(\bar{\alpha},\bar{\delta})$ rather than assuming a simple top hat. Although for simplicity we do not apply the pencil-beam approach in this Letter, it can be thought of as a natural interpolation between the counterpart and statistical cases.

To calculate the normalization term, $\beta(H_0)$, in the denominator of equation (2), we must account for selection effects in our measurement process. In general, the GW and EM data are both subject to selection effects in that we only detect GW and EM sources that are above some threshold, d_{GW}^{th} and d_{EM}^{th}, respectively. Accounting for these detection thresholds, the denominator of equation (2) is:

$$\beta(H_0) = \int_{d_{EM}>d_{EM}^{\text{th}}}\int_{d_{GW}>d_{GW}^{\text{th}}}\int^4 p(d_{GW},d_{EM},D_L,\alpha,\delta,z|H_0)\mathrm{d}D_L\mathrm{d}z\mathrm{d}\alpha\mathrm{d}\delta\mathrm{d}d_{GW}\mathrm{d}d_{EM} \tag{9}$$

We define:

$$P_{\text{det}}^{\text{GW}}(D_L,\alpha,\delta,z) \equiv \int_{d_{GW}>d_{GW}^{\text{th}}} p(d_{GW}|D_L,\alpha,\delta,z)\mathrm{d}d_{GW} \tag{10}$$

and similarly:

$$P_{\text{det}}^{\text{EM}}(z,\alpha,\delta) \equiv \int_{d_{EM}>d_{EM}^{\text{th}}} p(d_{EM}|z,\alpha,\delta)\mathrm{d}d_{EM} \tag{11}$$

With these definitions, equation (9) becomes:

$$\beta(H_0) = \int P_{\text{det}}^{\text{GW}}(\hat{D}_L(z,H_0),\alpha,\delta,z)P_{\text{det}}^{\text{EM}}(z,\alpha,\delta)p_0(z,\alpha,\delta)\mathrm{d}\alpha\mathrm{d}\delta\mathrm{d}z$$

Note that we have applied the same chain-rule factorization to the inner four integrals as in equation (3). It is clear that the normalization factor $\beta(H_0)$ depends on H_0, so it is crucial to include it in the likelihood. For the EM selection effects, we assume that we can detect all EM counterparts and host galaxies up to some maximum true redshift, $z_{\max}$. This is an over-simplification of the true EM selection effects, but is a reasonable assumption for the real-time galaxy catalogues that will be constructed during the EM follow-up to GW events. For example, at Advanced LIGO design sensitivity, 90% of $30M_\odot$–$30M_\odot$ BBH detections will be within 5 Gpc (and BNS detections will be within 0.3 Gpc)[43] (see http://gwc.rcc.uchicago.edu/). Furthermore, only the BBHs with the smallest localization volumes contribute to the H_0 constraints, and these will typically be within 400 Mpc. A galaxy with the same absolute magnitude as the Milky Way would have an apparent magnitude of <23 at typical $30M_\odot$–$30M_\odot$ BBH distances, or <17.5 for well localized BBHs, and <17 at typical BNS distances. Meanwhile, we expect kilonova counterparts to BNS mergers to have magnitudes of ≤21.7 on the first night, even at the farthest distances detectable by the HLVJI network at design sensitivity (assuming the optical counterpart to GW170817 is typical[13,44]). This is well within the magnitude limits of upcoming survey telescopes (for example, the Large Synoptic Survey Telescope), as well as within reach of current instruments (such as the Dark Energy Camera, the Subaru Hyper Suprime-Cam, the Zwicky Transient Facility and so on).

We assume that EM counterparts are detectable for binaries regardless of the binary inclination. Although the short γ-ray burst emission is expected to be beamed, the associated kilonovae are expected to emit isotropically. Furthermore, as GW170817 demonstrated, it is possible to identify a kilonova counterpart independently of the γ-ray burst. We note that since face-on/face-off binaries are louder (have higher SNR) than edge-on binaries in GWs, the population of detected binaries will show a preference for face-on/face-off; our analysis reproduces the expected inclination distribution among detected sources (see figure 4 of ref. [45]). Under these assumptions for the EM selection effects, we have:

$$P_{\text{det}}^{\text{EM}}(z,\alpha,\delta) \propto \mathcal{H}(z_{\max}-z) \tag{12}$$

where $\mathcal{H}$ is the Heaviside step function, and equation (9) reduces to:

$$\beta(H_0) = \int_0^{z_{\max}}\iint P_{\text{det}}^{\text{GW}}(\hat{D}_L(z,H_0),\alpha,\delta,z)p_0(z,\alpha,\delta)\mathrm{d}\alpha\mathrm{d}\delta\mathrm{d}z$$

Meanwhile, we assume that the galaxy distribution is approximately isotropic on large scales, so that the galaxy catalogue prior can be factored as:

$$p_0(z,\alpha,\delta) \approx p_0(z)p_0(\alpha,\delta) \tag{13}$$

and we approximate $p_0(\alpha,\delta)$ by a continuous, isotropic distribution on the sky. We note that this assumption would only introduce systematic errors if the galaxy distribution had strong correlations with the sky sensitivities of the detectors, which is not to be expected. Although GWs experience redshifting, the effect of redshift on the detectability of the GW data is small. We can therefore define:

$$P_{\text{det}}^{\text{GW}}(D_L) = \int P_{\text{det}}^{\text{GW}}(D_L,\alpha,\delta)p_0(\alpha,\delta)\mathrm{d}\alpha\mathrm{d}\delta \tag{14}$$

We assume a detection threshold for GW sources corresponding to a matched-filter network SNR $\rho_{\text{th}} = 12$, so that the detection probability, $P_{\text{det}}^{\text{GW}}(D_L)$, is the probability that a source at distance D_L will have SNR $\rho > 12$. Assuming a distribution of orientations, masses and spins among a population of sources, in addition to the assumed isotropic distribution on the sky, $p_0(\alpha,\delta)$, we calculate the fraction of sources that are detectable at a given distance, $P_{\text{det}}^{\text{GW}}(D_L)$. We assume an isotropic distribution of orientation angles, and fix spins to be zero. For simplicity, we assume a monochromatic mass distribution for each type of source. (For example, we take all BNS sources to be $1.4M_\odot$–$1.4M_\odot$.) We therefore have:

$$\beta(H_0) = \int_0^{z_{\max}} P_{\text{det}}^{\text{GW}}(D_L(z,H_0))p_0(z)\mathrm{d}z \tag{15}$$

We note that for GW sources in the local Universe ($D_L < 50$ Mpc):

$$\hat{D}_L(z, H_0) \approx \frac{cz}{H_0} \tag{16}$$

If we assume that the distribution of galaxies is uniform in comoving volume, then in the local universe, we can approximate:

$$p_0(z) \propto z^2 \tag{17}$$

With these approximations, assuming that EM selection effects are negligible ($z_{\max} \to \infty$), $\beta(H_0)$ is independent of the masses of the source, which determine the distance to which it can be detected. In fact, under these assumptions, $\beta(H_0)$ simplifies to:

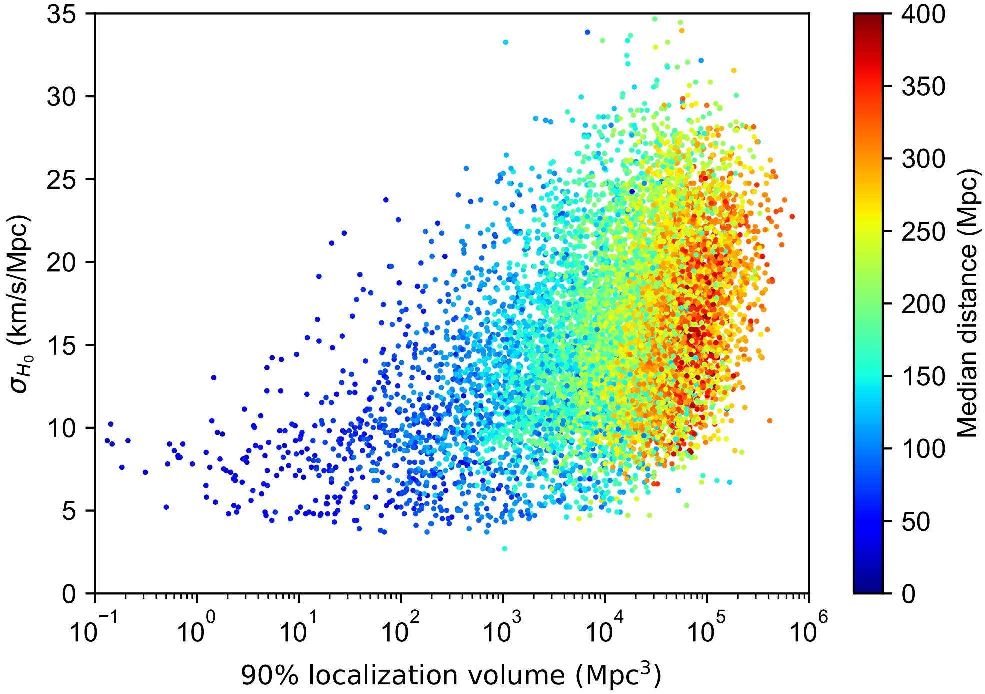
$$\beta(H_0) \propto H_0^3 \tag{18}$$

However, in general, we must account for cosmological deviations from equations (16) and (17), so we calculate $\beta(H_0)$ according to equation (15) throughout our analysis. We note that $\beta(H_0)$ is still only weakly dependent on the GW horizon and therefore on the (unknown) underlying mass distribution of GW sources. Nevertheless, the statistical framework described here can accommodate more complicated models of the GW source distribution and its effects on the detection probability (equation (14)).

Data availability

Source Data for Figs. 1, 2 and Extended Data Fig. 1 are provided with the online version of the paper. Other data that support the findings of this study are available from the corresponding author upon reasonable request.

28. Chen, H.-Y. & Holz, D. E. Finding the one: identifying the host galaxies of gravitational-wave sources. Preprint at https://arxiv.org/abs/1612.01471 (2016).
29. Abbott, B. P. et al. Prospects for observing and localizing gravitational-wave transients with advanced LIGO and advanced Virgo. *Liv. Rev. Rel*. **19**, 1 (2016).
30. Veitch, J. et al. Parameter estimation for compact binaries with ground-based gravitational-wave observations using the LALInference software library. *Phys. Rev. D* **91**, 042003 (2015).
31. Carrick, J., Turnbull, S. J., Lavaux, G. & Hudson, M. J. Cosmological parameters from the comparison of peculiar velocities with predictions from the 2M++ density field. *Mon. Not. R. Astron. Soc*. **450**, 317–332 (2015).
32. Scolnic, D. M. et al. The complete light-curve sample of spectroscopically confirmed type Ia supernovae from Pan-STARRS1 and cosmological constraints from the combined Pantheon sample. *Astrophys. J.* **859**, 101 (2018).
33. Del Pozzo, W. Inference of cosmological parameters from gravitational waves: applications to second generation interferometers. *Phys. Rev. D* **86**, 043011 (2012).
34. Fosalba, P., Crocce, M., Gaztañaga, E. & Castander, F. J. The MICE grand challenge lightcone simulation—I. Dark matter clustering. *Mon. Not. R. Astron. Soc.* **448**, 2987–3000 (2015).
35. Crocce, M., Castander, F. J., Gaztañaga, E., Fosalba, P. & Carretero, J. The MICE Grand Challenge lightcone simulation—II. Halo and galaxy catalogues. *Mon. Not. R. Astron. Soc*. **453**, 1513–1530 (2015).
36. Fosalba, P., Gaztañaga, E., Castander, F. J. & Crocce, M. The MICE Grand Challenge light-cone simulation—III. Galaxy lensing mocks from all-sky lensing maps. *Mon. Not. R. Astron. Soc*. **447**, 1319–1332 (2015).
37. Carretero, J. et al. CosmoHub and SciPIC: massive cosmological data analysis, distribution and generation using a Big Data platform. In *Proc. 2017 European Physical Society Conference on High Energy Physics* 488 (EPS-HEP, 2017).
38. Schechter, P. An analytic expression for the luminosity function for galaxies. *Astrophys. J*. **203**, 297–306 (1976).
39. Norberg, P. et al. The 2dF Galaxy Redshift Survey: the b_J-band galaxy luminosity function and survey selection function. *Mon. Not. R. Astron. Soc.* **336**, 907–931 (2002).
40. Liske, J., Lemon, D. J., Driver, S. P., Cross, N. J. G. & Couch, W. J. The Millennium Galaxy Catalogue: $16 \leq B_{MGC} < 24$ galaxy counts and the calibration of the local galaxy luminosity function. *Mon. Not. R. Astron. Soc*. **344**, 307–324 (2003).
41. González, R. E., Lares, M., Lambas, D. G. & Valotto, C. The faint-end of the galaxy luminosity function in groups. *Astron. Astrophys*. **445**, 51–58 (2006).
42. Kourkchi, E. & Tully, R. B. Galaxy groups within 3500 km s^{-1}. *Astrophys. J.* **843**, 16 (2017).
43. Chen, H.-Y. et al. Distance measures in gravitational-wave astrophysics and cosmology. Preprint at https://arxiv.org/abs/1709.08079 (2017).
44. Abbott, B. P. et al. Multi-messenger observations of a binary neutron star merger. *Astrophys. J*. **848**, 12 (2017).
45. Schutz, B. F. Networks of gravitational wave detectors and three figures of merit. *Class. Quant. Gravity* **28**, 125023 (2011).
46. Ade, P. A. R. et al. Planck 2015 results. XIII. Cosmological parameters. *Astron. Astrophys.* **594**, A13 (2016).



Extended Data Fig. 1 | H_0 uncertainty for BNS systems with identified counterparts and redshifts. Each point is the H_0 uncertainty σ_{H_0} from a simulated detection with the Advanced HLV network operating at design sensitivity, as a function of the 90% localization volume. The colours correspond to the median of the GW distance measurement. The lower limit to the precision of individual measurements of about 3 km s^{-1} Mpc^{-1} is due to the 'sweet spot' between peculiar velocities and distance uncertainties, as discussed in the text. We find that, in general, closer events have smaller localization volumes and lead to better constraints on H_0, although the closest events yield slightly worse constraints because of their larger fractional peculiar velocity uncertainties.

https://doi.org/10.1038/s41586-018-0559-3

Device-independent quantum random-number generation

Yang Liu[1,2], Qi Zhao[3], Ming-Han Li[1,2], Jian-Yu Guan[1,2], Yanbao Zhang[5], Bing Bai[1,2], Weijun Zhang[4], Wen-Zhao Liu[1,2], Cheng Wu[1,2], Xiao Yuan[1,2,3], Hao Li[4], W. J. Munro[5], Zhen Wang[4], Lixing You[4], Jun Zhang[1,2], Xiongfeng Ma[3]*, Jingyun Fan[1,2]*, Qiang Zhang[1,2]* & Jian-Wei Pan[1,2]*

Randomness is important for many information processing applications, including numerical modelling and cryptography[1,2]. Device-independent quantum random-number generation (DIQRNG)[3,4] based on the loophole-free violation of a Bell inequality produces genuine, unpredictable randomness without requiring any assumptions about the inner workings of the devices, and is therefore an ultimate goal in the field of quantum information science[5–7]. Previously reported experimental demonstrations of DIQRNG[8,9] were not provably secure against the most general adversaries or did not close the 'locality' loophole of the Bell test. Here we present DIQRNG that is secure against quantum and classical adversaries[10–12]. We use state-of-the-art quantum optical technology to create, modulate and detect entangled photon pairs, achieving an efficiency of more than 78 per cent from creation to detection at a distance of about 200 metres that greatly exceeds the threshold for closing the 'detection' loophole of the Bell test. By independently and randomly choosing the base settings for measuring the entangled photon pairs and by ensuring space-like separation between the measurement events, we also satisfy the no-signalling condition and close the 'locality' loophole of the Bell test, thus enabling the realization of the loophole-free violation of a Bell inequality. This, along with a high-voltage, high-repetition-rate Pockels cell modulation set-up, allows us to accumulate sufficient data in the experimental time to extract genuine quantum randomness that is secure against the most general adversaries. By applying a large (137.90 gigabits $\times$ 62.469 megabits) Toeplitz-matrix hashing technique, we obtain 6.2469×10^7 quantum-certified random bits in 96 hours with a total failure probability (of producing a random number that is not guaranteed to be perfectly secure) of less than 10^{-5}. Our demonstration is a crucial step towards transforming DIQRNG from a concept to a key aspect of practical applications that require high levels of security and thus genuine randomness[7]. Our work may also help to improve our understanding of the origin of randomness from a fundamental perspective.

The security, or unpredictability, of randomness generated by a device-independent quantum random-number generator can be assessed via the observation of the loophole-free violation of a Bell inequality. A Bell test involves two entangled particles, with each party choosing the measurement settings according to a random input and outputting a classical bit. To create a device-independent quantum random-number generator based on the violation of a Bell inequality, two sets of conditions must be fulfilled rigorously and simultaneously.

First, it is necessary in the experimental implementations of DIQRNG to detect entangled particles with high efficiency in order to close the detection loophole and to ensure the no-signalling condition—which requires that there is no information exchange between the preparation of input randomness and the preparation of entangled particles, or between the measurement setting at one detector and measurement outcome at the other—by space-like separating relevant events. Alternatively, proper shielding could be applied to prohibit communications between relevant events[7,8]; however, in practice, it is impossible to shield all of the known and unknown types of communication. Although recent progress in loophole-free tests of Bell inequalities[13–16] provides a way of realizing DIQRNG based on the violation of Bell inequalities, the implementation demands unprecedented detection efficiency and system stability. Therefore, DIQRNG remains a formidable challenge.

Second, an independent and identical distribution (i.i.d.) must not be assumed for the behaviour of the adversary, the most general quantum adversary should be considered in the security analysis and the production of random bits must occur at a non-vanishing rate and be noise-tolerant. In the i.i.d. scenario, assuming that the adversarial strategy follows a predetermined probability distribution, the security analysis is greatly simplified, even without considering any internal memory or time-dependent behaviour; however, the i.i.d. assumption in these theories fails in practice because, for example, the adversary may attack the system using the previous results in an adaptive, non-i.i.d. way. Although security against the most general quantum adversaries and without the i.i.d. assumption has been rigorously proved[12,17–21], a method for security analysis that is efficient for a non-infinite amount of data (and can therefore be tested experimentally) was demonstrated only very recently. A method for DIQRNG that does not use the i.i.d. assumption and that considers a general quantum adversary was proposed recently[12], based on the entropy accumulation theorem[21]. With this method, the rate of randomness generation approaches the value for the i.i.d. case in the limit of a large amount of data. We have previously presented an experimental demonstration of DIQRNG[22] that closed the detection loophole but did not consider space-like-separated events. Here we report fully functional DIQRNG, evidenced by rigorously satisfying the two sets of conditions discussed above in our experimental set-up and by accounting for general quantum adversaries in the security analysis. The device-independent quantum random-number generator that we demonstrate outputs genuinely, quantum-certified random bits at a rate of 181 bits s^{-1}—an important step towards practical applications.

Our realization of DIQRNG is based on a sequence of Bell-test experiments in the format of the Clauser–Horne–Shimony–Holt (CHSH) game[23]. We assume neither modelling of the physical apparatus nor any relation between different experimental trials. Time-dependent or memory-like effects may be present across experimental trials. We adopt the spot-checking protocol[10,12,24] for our experimental implementation. In experimental trial *i*, Alice and Bob, who are spatially separated, each receives a photon from an entangled pair. A classical

[1]National Laboratory for Physical Sciences at Microscale and Department of Modern Physics, University of Science and Technology of China, Hefei, China. [2]Shanghai Branch, CAS Center for Excellence and Synergetic Innovation Center in Quantum Information and Quantum Physics, University of Science and Technology of China, Shanghai, China. [3]Center for Quantum Information, Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. [4]State Key Laboratory of Functional Materials for Informatics, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China. [5]NTT Basic Research Laboratories and NTT Research Center for Theoretical Quantum Physics, NTT Corporation, Atsugi, Japan. *e-mail: xma@tsinghua.edu.cn; fanjy@ustc.edu.cn; qiangzh@ustc.edu.cn; pan@ustc.edu.cn

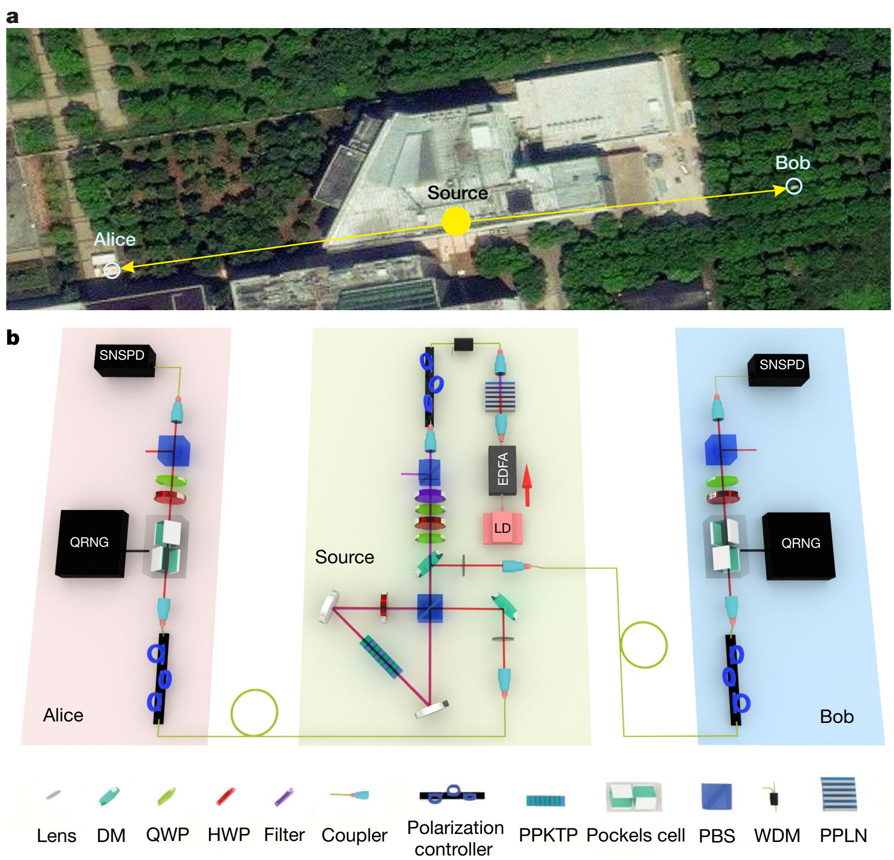


Fig. 1 | Schematics of the experiment. a, Top view of the experimental layout. Alice's and Bob's measurement stations are on the opposite sides of the source of entangled photon pairs, at distances of 93 ± 1 m and 90 ± 1 m, respectively, where the errors quoted are system errors. (We measure the distance by holding whiteboards and measuring the distance with a laser ranger; the error measured this way is estimated to be less than 1 m.) **b**, Middle, creation of pairs of entangled photons. Light pulses of 10-ns duration and 200-kHz frequency from a 1,560-nm laser diode (LD) are amplified by an erbium-doped fibre amplifier (EDFA) and frequency-doubled in an in-line periodically poled lithium niobate crystal (PPLN); the red arrow indicates the direction of the light. The resultant 780-nm light pulses are focused into a periodically poled potassium titanyl phosphate crystal (PPKTP) in a Sagnac loop to generate polarization-entangled photon pairs. A half-wave plate (HWP) and two quarter-wave plates (QWPs) are used to control the relative amplitude and phase in the polarization-entangled two-photon state that is created. Left and right, measuring the polarization of single photons. The single photons exit the fibre, experience the polarization-state measurement in free space and are collected into single-mode optical fibres to be detected by superconducting nanowire single-photon detectors (SNSPDs). The apparatus that is used to perform the measurement consists of a Pockels cell, a quarter-wave plate, a half-wave plate and a polarizing beam splitter (PBS). Quantum random-number generators (QRNGs) output random bits, triggering the Pockels cell to switch between two polarization orientations. DM, dichroic mirror; WDM, wavelength-division multiplexer. Underlying map in **a** from Google, DigitalGlobe.

bit $t_i = 0$ or $t_i = 1$ is generated with probability $1 - q$ or q, respectively. If $t_i = 0$, then the trial is a 'generation' trial, with fixed inputs for Alice and Bob, $x_i = 0$ and $y_i = 0$, respectively. If $t_i = 1$, then the trial is a 'test' trial to test against adversaries. In our experiment, we set all trials to be test trials by choosing $q = 0$. In each test trial, Alice and Bob each receives a bit from a quantum random-number generator, $x_i, y_i \in \{0, 1\}$, as an input that determines their measurement setting. Alice's (Bob's) measurement setting is not affected by Bob's (Alice's) measurement setting or measurement outcome, and is independent of the entanglement creation at the source. Hence, the experiment satisfies the no-signalling condition. We assume that the two random inputs, x_i and y_i, are created independently of the rest of the experiment, and that their creation is i.i.d. for all of the n trials. The corresponding measurement outcomes are $a_i, b_i \in \{0, 1\}$. We assign a CHSH game value of $J_i = 1$ if $a_i \oplus b_i = x_i y_i$ and of $J_i = 0$ otherwise.

Considering uniform inputs for all n experimental trials (so that the probability of selecting any $x_i, y_i \in \{0, 1\}$ is 1/4), the CHSH game value $\bar{J}$ over all n experimental trials is

$$\bar{J} = \frac{1}{n}\sum_{i=1}^{n} J_i - 3/4$$

The experiment is subject to various possible loss mechanisms. We require that the photon loss is low enough to close the detection loophole. Any adversarial strategies based on local hidden-variable models yield $\bar{J} \le 0$. Therefore, $\bar{J} > 0$ indicates that the measurement outcomes cannot be pre-determined and therefore represent genuine, unpredictable quantum randomness.

The amount of unpredictable randomness that can be extracted in the presence of the quantum adversary system E is quantified by the smooth min-entropy[12]:

$$H_{\min}^{\varepsilon_s}(\mathbf{AB}|\mathbf{XY}E) \ge nR_{\mathrm{opt}}(\varepsilon_s, \varepsilon_{\mathrm{EA}}, \omega_{\exp})$$

with smoothing parameter ε_s, expected CHSH game value $\omega_{\exp}$ and failure probability for the entropy accumulation protocol $\varepsilon_{\mathrm{EA}}$. **X** and **Y** denote the input sequences for Alice and Bob, respectively, and **A** and **B** the corresponding output sequences. The lower bound of the generation rate, $R_{\mathrm{opt}}(\varepsilon_s, \varepsilon_{\mathrm{EA}}, \omega_{\exp})$, is used as the theoretical amount of randomness on average for each trial. See Supplementary Information section I.B for details. By using a Toeplitz-matrix hashing extractor with a size of $n \times [H_{\min}^{\varepsilon_s}(\mathbf{AB}|\mathbf{XY}E) - t_e]$, where t_e is the number of bits sacrificed to minimize the information that an adversary may acquire (a parameter that is relevant to the failure probability of the extractor),

we extract $H_{\min}^{\varepsilon_s}(\mathbf{AB}|\mathbf{XY}E) - t_e$ random bits with genuine unpredictability from the raw data obtained in the n experimental trials. The random bits are $\varepsilon_s + \varepsilon_{EA} + 2^{-t_e}$ close to a uniform distribution, where 2^{-t_e} is the failure probability in extraction, with $t_e = 100$.

Although we do not assume the inner workings of the devices, we still require a few assumptions in our experimental implementation: (1) that the devices and adversaries behave according to the laws of either classical or quantum mechanics; (2) that Alice's and Bob's devices are located in one secure laboratory so that the adversaries cannot access their measurement outcomes; (3) that Alice and Bob each receives a sequence of uniform random bits to determine their measurement setting from an independent, trusted source; and (4) that Alice and Bob each has a trusted classical post-processing unit to extract the final random bits. In general, device-independent protocols (in particular for quantum key distribution) assume that Alice's and Bob's devices are located in two secure laboratories and in between them there is a classical authentication channel. In these cases, the publicly announced information may be leaked to the adversaries. The security is then compromised by reusing the devices. Any attack based on this premise is known as a memory attack[25]. With the above assumptions and by ensuring no information leakage, our implementation is secure against memory attacks. In our experiment, quantum random numbers are required as inputs to switch the measurement basis in the Bell inequality. In this sense, our experiment can be seen as a type of randomness expansion, which generates more randomness from a random seed. An interesting future direction of research would be to create nearly perfect random numbers from weak randomness, which may require more than one set of Bell-test equipment.

Our experimental implementation is depicted in Fig. 1. We create entangled photon pairs at a wavelength of 1,560 nm using spontaneous parametric downconversion in a Sagnac interferometer (see Methods). We then distribute the two photons of a pair in opposite directions to Alice's and Bob's measurement stations, which are at distances of 93 m and 90 m from the source, respectively. A detailed space–time analysis (Fig. 2) shows that the relevant events in the experiment are space-like-separated (Supplementary Information section II.E). We obtain an overall efficiency from the creation to the detection of the entangled photons of 78.8% ± 1.9% for Alice and 78.5% ± 1.5% for Bob[26] (where the errors quoted are one standard deviation), surpassing the threshold to close the detection loophole (Supplementary Information section II.C).

To achieve the maximum violation of the Bell inequality[27], we create a non-maximally polarization-entangled two-photon state, $\cos(22.05°)|HV\rangle + \sin(22.05°)|VH\rangle$ and choose the measurement settings to be $-83.5°$ (for $x_i = 0$) or $-119.4°$ (for $x_i = 1$) for Alice and $6.5°$ (for $y_i = 0$) or $-29.4°$ (for $y_i = 1$) for Bob when measuring the polarization state of the entangled photons. The measurement settings are selected randomly by the quantum random-number generators in each experimental trial. The two quantum random-number generators are based on vacuum noise fluctuation (Supplementary Information section II.A).

Our system is now robust against noise, which allows us to complete $n = 6.895 \times 10^{10}$ experimental trials in 95.77 experimental hours, operating continuously. To quantify the significance of our experimental results, we perform a hypothesis test of local realism. The null hypothesis is that the experimental results are explainable by local realism under the assumption that the input distribution at each trial is uniform. The evidence against local realism, under the above assumption, is quantified by a statistical P value computed using a test statistic. The P value is the maximum probability according to local realism that the statistic takes a value as extreme as the observed one. Hence, small P values imply strong evidence against local realism. We apply the prediction-based ratio (PBR) method of analysis[28] to design the test statistic and compute an upper bound for the P value. The PBR analysis provides valid upper bounds for P values without assuming the i.i.d. condition. The upper bound that is returned after the whole experiment is $p_{LR} = 10^{-204,792}$, indicating a strong rejection of local

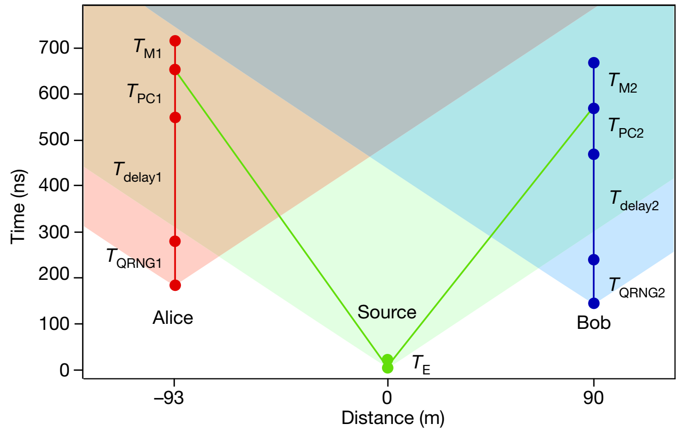


Fig. 2 | Space–time diagram for the experimental design. $T_E = 10$ ns is the time taken to generate a pair of entangled photons. $T_{QRNG1,2}$ are the times required to generate random bits to switch the Pockels cells. $T_{delay1,2}$ are the lengths of time between the random bits being generated and delivered to the Pockels cells. $T_{PC1,2}$ are the waiting times for the Pockels cells to be ready to perform state measurements after receiving the random bits. $T_{M1,2}$ are the times taken by the superconducting nanowire single-photon detectors to output electronic signals. For $T_{QRNG1} = T_{QRNG2} = 96$ ns, $T_{delay1} = 270$ ns, $T_{delay2} = 230$ ns, $T_{PC1} = 112$ ns, $T_{PC2} = 100$ ns, $T_{M1} = 50$ ns and $T_{M2} = 100$ ns, we place Alice's and Bob's measurement stations on opposite sides of the source at distances of 93 m and 90 m, respectively. The effective optical length between Alice's (Bob's) station and the source is 132 m (119 m). This arrangement ensures no signalling between relevant events in the experiment. The shaded areas are the future light cones for the source, Alice and Bob.

hidden-variable models (see Supplementary Information III.C). With the PBR method, we also test the null hypothesis that the experimental results satisfy the no-signalling condition under the assumption that the input distribution at each trial is uniform. In this case, we obtain an upper bound of $p_{NS} = 1$, indicating no evidence of anomalous signalling in the experiment (see Supplementary Information section III.B).

We compute the CHSH game value J over n experimental trials to be $\bar{J} = 2.757 \times 10^{-4}$. By setting the expected CHSH game value to that measured in the experiment, $\omega_{exp} = 2.757 \times 10^{-4}$, and assuming that $\varepsilon_s = \varepsilon_{EA} = \sqrt{1/n} = 3.8 \times 10^{-6}$ and that the width of the statistical confidence interval for the estimate of the Bell violation is $\delta_{est} = \sqrt{10/n} = 1.2042 \times 10^{-5}$, we find a total failure probability of $\varepsilon_s + \varepsilon_{EA} + 2^{-t_e} < 1 \times 10^{-5}$. After developing a computing technique for fast Toeplitz-matrix multiplication (Supplementary Information section II.H) that allows us to apply an 137.90 Gb × 62.469 Mb Toeplitz-matrix hashing, we obtain 6.2469×10^7 genuinely, quantum-certified random bits, corresponding to a rate of 181.2 bits s^{-1}, with a total failure probability of 10^{-5}. The stream of random bits passes the National Institute of Standards and Technology (NIST) statistic test suite (Supplementary Information section III.A). As shown in Fig. 3, the amount of randomness generated with our experimental set-up is 56.9% of the optimal value in the i.i.d. case for $n = 6.895 \times 10^{10}$, and asymptotically approaches the optimal value with an increasing number of experimental trials. In the inset of Fig. 3 we plot the randomness production as a function of time, demonstrating the robustness of the system.

In conclusion, we report here the realization of DIQRNG that is secure against the most general quantum adversaries and outputs 181 quantum-certified random bits per second. A next step will be to improve the violation of the Bell inequality and the stability of the system to achieve higher production rates of quantum-certified random bits for practical applications that require high levels of security. We anticipate that our work will be helpful in topics such as randomness amplification[29], the minimum assumption necessary for randomness generation, and fundamental problems relating to the understanding of non-locality, entanglement and randomness[7].

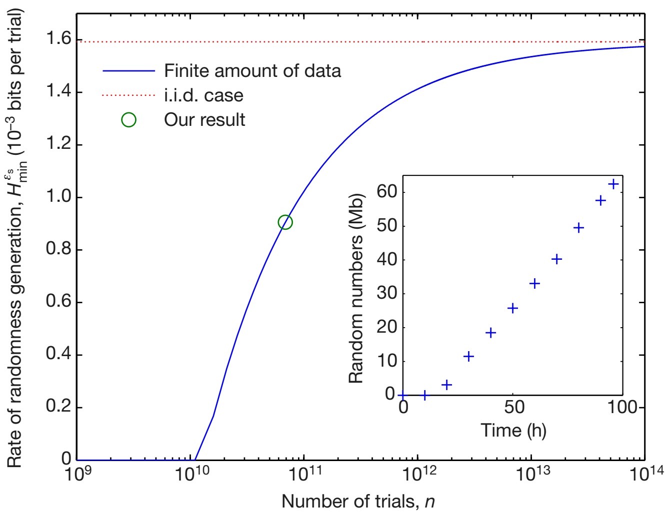


Fig. 3 | Randomness generation versus number of trials. The solid blue curve and dashed red curve show the theoretical rate of randomness generation $H_{\min}^{\varepsilon_s}$ versus the number of trials n for a finite amount of data (up to 10^{14}) and the i.i.d. case, respectively. The green circle denotes our experimental result. The inset shows the number of random numbers generated versus time in our experiment. We set $\omega_{\text{exp}} = 2.757 \times 10^{-4}$, $\varepsilon_s = \varepsilon_{\text{EA}} = 1/\sqrt{n}$ and $\delta_{\text{est}} = \sqrt{10/n}$ for the finite data rate.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0559-3.

Received: 7 February 2018; Accepted: 23 August 2018;
Published online 19 September 2018.

1. Shannon, C. E. Communication theory of secrecy systems. *Bell Labs Tech. J.* **28**, 656–715 (1949).
2. Metropolis, N. & Ulam, S. The Monte Carlo method. *J. Am. Stat. Assoc.* **44**, 335–341 (1949).
3. Colbeck, R. *Quantum and Relativistic Protocols for Secure Multi-party Computation*. PhD thesis, Cambridge Univ. (2009).
4. Mayers, D. & Yao, A. Quantum cryptography with imperfect apparatus. In *Proc. 39th Annual Symposium on Foundations of Computer Science* (ed. Motwani, R.) 503–509 (IEEE, 1998).
5. Ma, X., Yuan, X., Cao, Z., Qi, B. & Zhang, Z. Quantum random number generation. *npj Quantum Inf.* **2**, 16021 (2016).
6. Herrero-Collantes, M. & Garcia-Escartin, J. C. Quantum random number generators. *Rev. Mod. Phys.* **89**, 015004 (2017).
7. Acín, A. & Masanes, L. Certified randomness in quantum physics. *Nature* **540**, 213–219 (2016).
8. Pironio, S. et al. Random numbers certified by Bell's theorem. *Nature* **464**, 1021–1024 (2010).
9. Bierhorst, P. et al. Experimentally generated randomness certified by the impossibility of superluminal signals. *Nature* **556**, 223–226 (2018).
10. Miller, C. A. & Shi, Y. Universal security for randomness expansion from the spot-checking protocol. *SIAM J. Comput.* **46**, 1304–1335 (2017).
11. Vazirani, U. V. & Vidick, T. Certifiable quantum dice - or, testable exponential randomness expansion. Preprint at https://arxiv.org/abs/1111.6054 (2011).
12. Arnon-Friedman, R., Renner, R. & Vidick, T. Simple and tight device-independent security proofs. Preprint at https://arxiv.org/abs/1607.01797 (2016).
13. Hensen, B. et al. Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature* **526**, 682–686 (2015).
14. Shalm, L. K. et al. Strong loophole-free test of local realism. *Phys. Rev. Lett.* **115**, 250402 (2015).
15. Giustina, M. et al. Significant-loophole-free test of Bell's theorem with entangled photons. *Phys. Rev. Lett.* **115**, 250401 (2015).
16. Rosenfeld, W. et al. Event-ready Bell test using entangled atoms simultaneously closing detection and locality loopholes. *Phys. Rev. Lett.* **119**, 010402 (2017).
17. Vazirani, U. & Vidick, T. Fully device-independent quantum key distribution. *Phys. Rev. Lett.* **113**, 140501 (2014).
18. Miller, C. A. & Shi, Y. Robust protocols for securely expanding randomness and distributing keys using untrusted quantum devices. *J. ACM* **63**, 33 (2016).
19. Chung, K.-M., Shi, Y. & Wu, X. Physical randomness extractors: generating random numbers with minimal assumptions. Preprint at https://arxiv.org/abs/1402.4797 (2014).
20. Coudron, M. & Yuen, H. Infinite randomness expansion with a constant number of devices. In *Proc. 46th Annual ACM Symposium on Theory of Computing* (ed. Shmoys, D.) 427–436 (ACM, 2014).
21. Dupuis, F., Fawzi, O. & Renner, R. Entropy accumulation. Preprint at https://arxiv.org/abs/1607.01796 (2016).
22. Liu, Y. et al. High-speed device-independent quantum random number generation without a detection loophole. *Phys. Rev. Lett.* **120**, 010503 (2018).
23. Clauser, J. F., Horne, M. A., Shimony, A. & Holt, R. A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* **23**, 880–884 (1969).
24. Coudron, M., Vidick, T. & Yuen, H. Robust randomness amplifiers: upper and lower bounds. In *Proc. APPROX 2013: Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques* (eds Raghavendra, P. et al.) 468–483 (Springer, 2013).
25. Barrett, J., Colbeck, R. & Kent, A. Memory attacks on device-independent quantum cryptography. *Phys. Rev. Lett.* **110**, 010503 (2013).
26. Pereira, M. D. C. et al. Demonstrating highly symmetric single-mode, single-photon heralding efficiency in spontaneous parametric downconversion. *Opt. Lett.* **38**, 1609–1611 (2013).
27. Eberhard, P. H. Background level and counter efficiencies required for a loophole-free Einstein-Podolsky-Rosen experiment. *Phys. Rev. A* **47**, R747–R750 (1993).
28. Zhang, Y., Glancy, S. & Knill, E. Asymptotically optimal data analysis for rejecting local realism. *Phys. Rev. A* **84**, 062118 (2011).
29. Kessler, M. & Arnon-Friedman, R. Device-independent randomness amplification and privatization. Preprint at https://arxiv.org/abs/1705.04148 (2017).

Acknowledgements We thank S.-R. Zhao, Y.-H. Li, L.-K. Chen and R. Jin for experimental assistance, J. Zhong and S.-C. Shi for low-temperature system maintenance, and T. Peng, Y. Cao, C.-Z. Peng and Y.-A. Chen for discussions. This work was supported by the National Key R&D Program of China (2017YFA0303900, 2017YFA0304000), the National Natural Science Foundation of China, the Chinese Academy of Sciences and the Anhui Initiative in Quantum Information Technologies.

Reviewer information *Nature* thanks R. Colbeck and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions X.M., J.F., Q.Z. and J.-W.P. conceived the research. Y.L., X.M., J. F., Q.Z. and J.-W.P. designed the experiment. Y.L., M.-H.L. and C.W. designed and implemented the source of entangled photon pairs. W.-Z.L. and J.-Y.G. designed the data acquisition software. W.Z., H.L., Z.W. and L.Y. fabricated and characterized the superconducting nanowire single-photon detector. B.B. and J.Z. designed the quantum random-number generators for the measurement setting choices. Q. Zhao, X.Y. and X.M. performed the protocol analysis, numerical modelling and randomness extraction. Y.Z. and W.J.M. performed the hypothesis tests. All authors contributed to the experimental realization, data analysis and manuscript preparation. J.-W.P. supervised the project.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0559-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to X.M. or J.F. or Q.Z. or J.-W.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

No statistical methods were used to predetermine sample size.

Entanglement creation, distribution, and detection at low loss. The experimental layout is depicted in Fig. 1. A 1,560-nm laser outputs 10-ns laser pulses periodically at a repetition rate of 200 kHz. The pulses are amplified by an erbium-doped fibre amplifier and frequency-doubled in an in-line periodically poled lithium niobate crystal. After removing the residual 1,560-nm light with a wavelength-division multiplexer and spectral filters, the 780-nm light pulses that are generated are focused into a periodically poled potassium titanyl phosphate (PPKTP) crystal with a poling period of 46.5 μm in a Sagnac loop to generate polarization-entangled photon pairs via a type-II spontaneous parametric downconversion process.

We optimize the efficiency to couple the entangled photons that are created into a single-mode optical fibre by setting their beam waists to the theoretical optimized size with respect to that of the pump beam[26,30,31]. For the pump light exiting a 780HP single-mode fibre, which has a mode field diameter of 5 μm, we focus it with a focal length of $f = 8$ mm aspherical lens to the centre of the crystal at a distance of 70 cm, with a measured beam waist of 180 μm and a beam quality factor of $M^2 = 1.05$. The entangled photons are collected into a SMF28e single-mode optical fibre, which has a mode field diameter of 10.4 μm. With a $f = 11$ mm aspherical lens and a $f = 175$ mm spherical lens, we set the beam waist to be 85 μm at the centre of the PPKTP crystal. The spherical lens is at a distance of 19 cm from the aspherical lens and 45 cm from the PPKTP crystal.

A half-wave plate and two quarter-wave plates in the beam path of the pump light are used to control the relative amplitude and phase in the polarization-entangled two-photon state. The residual 780-nm pump light is removed by dichroic mirrors. The entanglement source is placed on a 1 m × 1 m breadboard, with the ambient temperature stabilized to be within ±1 °C to improve the stability of the system.

The entangled photons are sent in opposite directions to two remote measurement stations that are 93 ± 1 m and 90 ± 1 m away, ensuring space-like separation between the event of entanglement creation in the source and the event of choosing the measurement settings at the measurement stations, and between the event of choosing the measurement setting at one station and the events of choosing the measurement setting and outputting the outcomes at the other station (Supplementary Information section II.E).

At the measurement station, the single photons exit the fibre, pass the Pockels cell, a quarter-wave plate and a half-wave plate, and a polarizing beam splitter, and are coupled into the single-mode optical fibre to be detected by the superconducting nanowire single-photon detectors[32]. The Pockels cell is switched to set the base for the measurement of the single-photon polarization upon receiving a bit from a quantum random-number generator. A time–digital converter is used to time-tag the events for random-number generation, single-photon detection and synchronization signal.

We measure the overall efficiency from the creation to the detection of single photons to be 78.8% ± 1.9% for Alice and 78.5% ± 1.5% for Bob, surpassing the threshold to close the detection loophole. The loss is mainly due to the limited efficiency: 94% in collecting the photon pairs that are created into single-mode optical fibre, and 92% for the superconducting nanowire single-photon detectors (Supplementary Information section II.C).

Data availability

The data that support the findings of this study are available from the corresponding authors on reasonable request. Source Data for Fig. 3 is provided with the online version of the paper.

30. Bennink, R. Optimal collinear Gaussian beams for spontaneous parametric down-conversion. *Phys. Rev. A* **81**, 053805 (2010).
31. Dixon, P. B. et al. Heralding efficiency and correlated-mode coupling of near-IR fiber-coupled photon pairs. *Phys. Rev. A* **90**, 043804 (2014).
32. Zhang, W. et al. NbN superconducting nanowire single photon detector with efficiency over 90% at 1550 nm wavelength operational at compact cryocooler temperature. *Sci. China Phys. Mechan. Astron.* **60**, 120314 (2017).

https://doi.org/10.1038/s41586-018-0601-5

Exciton–polariton topological insulator

S. Klembt[1,5]*, T. H. Harder[1,5], O. A. Egorov[1], K. Winkler[1], R. Ge[2], M. A. Bandres[3], M. Emmerling[1], L. Worschech[1], T. C. H. Liew[2], M. Segev[3], C. Schneider[1] & S. Höfling[1,4]*

Topological insulators—materials that are insulating in the bulk but allow electrons to flow on their surface—are striking examples of materials in which topological invariants are manifested in robustness against perturbations such as defects and disorder[1]. Their most prominent feature is the emergence of edge states at the boundary between areas with different topological properties. The observable physical effect is unidirectional robust transport of these edge states. Topological insulators were originally observed in the integer quantum Hall effect[2] (in which conductance is quantized in a strong magnetic field) and subsequently suggested[3–5] and observed[6] to exist without a magnetic field, by virtue of other effects such as strong spin–orbit interaction. These were systems of correlated electrons. During the past decade, the concepts of topological physics have been introduced into other fields, including microwaves[7,8], photonic systems[9,10], cold atoms[11,12], acoustics[13,14] and even mechanics[15]. Recently, topological insulators were suggested to be possible in exciton-polariton systems[16–18] organized as honeycomb (graphene-like) lattices, under the influence of a magnetic field. Exciton-polaritons are part-light, part-matter quasiparticles that emerge from strong coupling of quantum-well excitons and cavity photons[19]. Accordingly, the predicted topological effects differ from all those demonstrated thus far. Here we demonstrate experimentally an exciton-polariton topological insulator. Our lattice of coupled semiconductor microcavities is excited non-resonantly by a laser, and an applied magnetic field leads to the unidirectional flow of a polariton wavepacket around the edge of the array. This chiral edge mode is populated by a polariton condensation mechanism. We use scanning imaging techniques in real space and Fourier space to measure photoluminescence and thus visualize the mode as it propagates. We demonstrate that the topological edge mode goes around defects, and that its propagation direction can be reversed by inverting the applied magnetic field. Our exciton-polariton topological insulator paves the way for topological phenomena that involve light–matter interaction, amplification and the interaction of exciton-polaritons as a nonlinear many-body system.

Microcavity exciton-polaritons (sometimes simply called polaritons hereafter) are composite bosons originating from the strong coupling of quantum-well excitons to microcavity photons. While the excitonic fraction provides strong nonlinearity, the photonic part results in a low effective mass, allowing the formation of a Bose–Einstein condensate that is dissipative in nature and can be driven by a laser beam[20]. Polaritons have thus been described as “quantum fluids of light”[21]. For the epitaxially well-controlled III–V semiconductor system, several techniques are available to micropattern cavities and thus precisely engineer the potential landscapes of polaritons[22]. With recent advances bringing topological effects into the realms of photonics and polaritonics[7–10,23], several ways to realize topological edge propagation with polaritons have been suggested[16–18], with honeycomb geometries (‘artificial graphene’[24]) being of particular interest to realize a Chern band insulator with Chern number $C=2$ (refs [17,18]). Polariton honeycomb lattices have been found to support Dirac-cone dispersions[25] as well as edge modes[26] inherited from their graphene origin[27,28]. Here, we take the next step and create a topological Chern insulator in a symbiotic part-light, part-matter system: the system of exciton-polaritons. Our experiments are based on the proposals[17,18] of a honeycomb potential landscape for exciton-polaritons, with its time-reversal symmetry broken by an applied magnetic field.

Let us first introduce the underlying physics and show its features in numerical simulations with a realistic set of sample parameters. The experiments display reliable condensation of polaritons in the vicinity of the Dirac cones and the creation of topological edge modes by applying a magnetic field. Hence, we consider the injection of polaritons into the topological gap and focus on the inherent properties of the chiral edge modes.

A general schematic of the experiment is presented in Fig. 1a. Figure 1b, c depicts the calculated dispersion relation of the honeycomb structure in the direction $\Gamma \rightarrow$ K, connecting two Dirac cones (K and K′), without and with an applied magnetic field, respectively. The effective spin–orbit coupling of polaritons, induced by transverse electric/transverse magnetic (TE–TM) mode splitting, breaks the polarization-related symmetry, and thus each Dirac point transforms into four inverted parabolas[18]. Whereas the spin–orbit interaction is extremely small in real graphene[29], ‘polariton graphene’ offers the possibility of making the effective spin–orbit coupling large enough to open a sizable gap in a magnetic field. Without a magnetic field, the two central parabolas touch each other at the Dirac cones of the underlying honeycomb lattice (Fig. 1b). The degeneracy between the states in the crossing points can be lifted in the presence of a magnetic field and a finite Zeeman splitting. As a consequence, an energy gap forms near the Dirac cones (Fig. 1c). It is worth mentioning that the Dirac cones K and K′ are not equivalent. At the K (K′) point, the ‘valence’ band is formed from the B (A) pillars, and the ‘conduction’ band is formed from the A (B) pillars[18]. The reversed order of the bands in the basis of the sublattices signifies that the gap is topologically non-trivial[1].

The interplay of an external magnetic field and the effective spin–orbit coupling (TE–TM mode splitting) results in non-zero Berry connections around the K and K′ points, contributing to the total band Chern number $C=\pm 2$ (refs [17,18]). As a consequence, the honeycomb structure supports one-way propagating edge states for the energies within the topological gap. Figure 1c demonstrates the results of a band structure calculation combined with the dispersion of the edge states localized at the zigzag edge of the honeycomb structure. The propagation direction of these edge states is related to the direction of the external magnetic field: the polariton edge current is either left-moving (L, yellow) or right-moving (R, red), depicted in Fig. 1a, depending on the sign of the magnetic field. Figure 1e depicts the corresponding calculated edge mode.

To illustrate the existence and robustness of the topologically protected one-way edge states, we simulate the evolution of a wavepacket excited locally (Fig. 1f, g; red circle) at the zigzag edge of the honeycomb structure. Figure 1f, g shows that the launched wavepacket starts

[1]Technische Physik and Wilhelm-Conrad-Röntgen-Research Center for Complex Material Systems, Universität Würzburg, Würzburg, Germany. [2]Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore, Singapore. [3]Physics Department and Solid State Institute, Technion, Haifa, Israel. [4]SUPA, School of Physics and Astronomy, University of St Andrews, St Andrews, UK. [5]These authors contributed equally: S. Klembt, T. H. Harder. *e-mail: sebastian.klembt@physik.uni-wuerzburg.de; sven.hoefling@physik.uni-wuerzburg.de

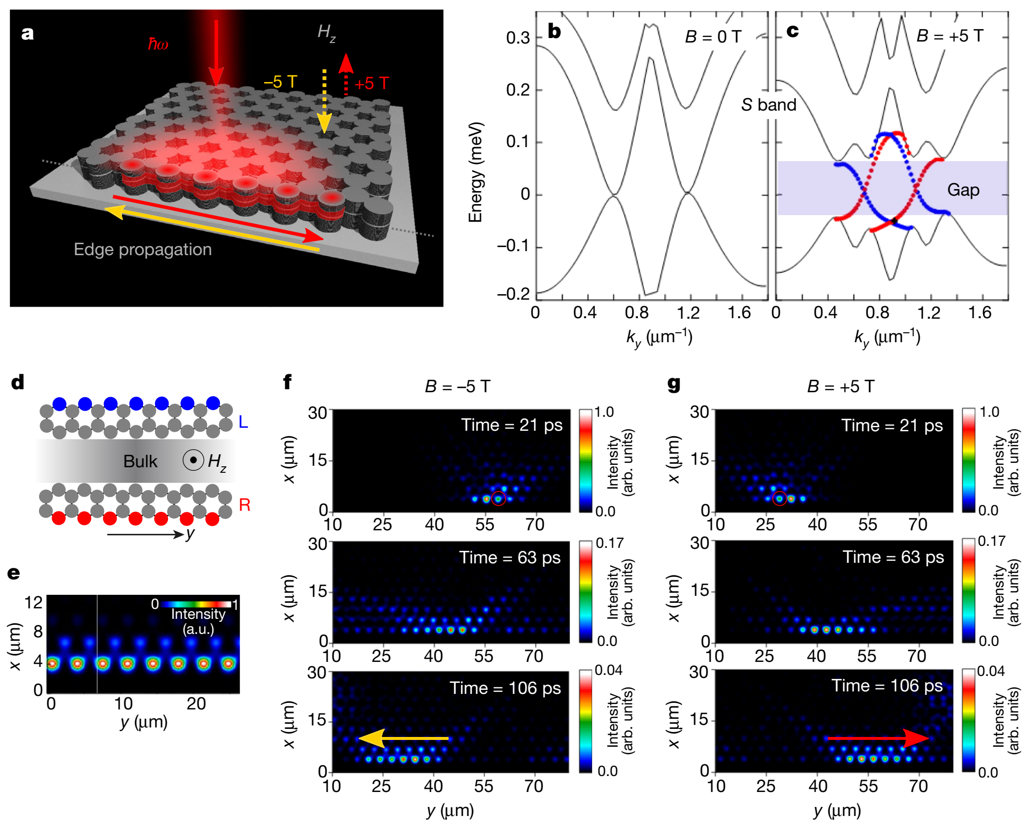


Fig. 1 | Experimental scheme, calculated band structure and simulated dynamics of the topological polariton edge modes. **a**, Schematic of the non-resonant laser excitation of left-moving (yellow) and right-moving (red) chiral topological polariton edge modes in a magnetic field H_z. **b**, **c**, Trivial and topological band structures of the polariton honeycomb lattice for zero Zeeman splitting (**b**) and with Zeeman splitting of $\Delta_{\mathrm{eff}} = 0.2$ meV, induced by the external magnetic field (**c**). One-way topological edge modes are represented by red (right-moving) and blue (left-moving) data points within the topological gap. **d**, **e**, Schematic (**d**) and calculated (**e**) intensity profiles of the edge modes. **f**, **g**, Calculated propagation dynamics of edge modes injected coherently (red circles) into the topological gap: left-moving time sequence (yellow) for negative Zeeman splitting $\Delta_{\mathrm{B}} = -0.8$ meV (**f**) and right-moving propagation (red) for positive Zeeman splitting $\Delta_{\mathrm{B}} = +0.8$ meV (**g**). arb. units, arbitrary units.

to propagate left or right along the edge, depending on the polarity of the magnetic field, whereas in the absence of magnetic field, the launched wavepacket remains at the excitation point. Note that the overall intensity decreases with propagation, as the model takes into account a realistic polariton lifetime of about 35 ps (see Methods for details). The simulations reveal that the chiral edge mode is topologically protected, as it propagates along the 90° corner and is able to pass a point-like defect without scattering (see Methods).

Having established the features that we expect to observe in an exciton-polariton topological insulator, we describe the experimental platform. To realize a honeycomb potential lattice, we fabricate a planar microcavity containing three $In_{0.04}Ga_{0.96}As$ quantum wells in a λ-cavity, sandwiched between two GaAs/AlGaAs distributed Bragg reflectors (DBRs) (Fig. 2a). Subsequently, we use electron beam lithography to define the honeycomb lattice, using micropillars with a diameter of $d = 2.0$ μm and a pillar-to-pillar overlap of $v = a/d = 0.85$, where a denotes the centre-to-centre distance between neighbouring pillars. Finally, the upper DBR is etched in such a way that only two DBR pairs of the top DBR remain, so as not to damage the active quantum-well region (Fig. 2b, c) (see Methods). When these sites are arranged in a two-dimensional lattice, the discrete pillar modes hybridize because of their proximity to one another and form a polaritonic band structure. The honeycomb lattice is characterized by a two-element base in real space (Fig. 2d) and two non-equivalent K and K′ points supporting Dirac cones in the first Brillouin zone (Fig. 2e), as well known from graphene. Figure 2f depicts characterization of the polariton honeycomb lattice in the linear regime, using non-resonant laser excitation. The Fourier-space energy-resolved photoluminescence of the investigated lattice is imaged in the k_y direction and scanned in the k_x direction. The blue data points are fitted to the measured dispersions, accurately revealing the six Dirac cones at the K and K′ points. The results of the corresponding tight-binding model (see Methods for details) are plotted in red and brown, agreeing well with the experimental data.

Next, we describe the experiments conducted on the honeycomb exciton-polariton lattice under an external magnetic field, aiming to find topologically non-trivial edge states. To be able to observe a bandgap that opens at the Dirac points, the Zeeman splitting and the TE–TM splitting at the Dirac points need to be sufficiently large compared with the photoluminescence linewidth. This implies that the polaritons need to have a sufficient excitonic part. Therefore, we select a lattice at a moderate negative exciton-to-photon detuning of $\delta = -11.5$ meV as summarized in the Methods.

To assess the size of the bandgap, we apply polarization-resolved spectroscopy, making use of a $\lambda/4$ polarization series at an external magnetic field of $B = 0$ T and 5 T. At an external magnetic field of $B = +5$ T, the Hopfield coefficients at the Dirac points yield a photonic fraction of $|C|^2 = 0.96$ and an excitonic fraction of $|X|^2 = 0.04$. Furthermore, the TE–TM splitting at the wavevector of the Dirac point (that is, $|k_{\mathrm{D}}| \approx 0.77\pi/a$) can be estimated to be 400 μeV for the photons, resulting in an effective TE–TM splitting for the exciton-polaritons of about 384 μeV ($\beta_{\mathrm{eff}} \approx 263$ μeV μm^2; see Methods). The Zeeman splitting of the excitonic mode is determined to be about 540 μeV, leading to an effective Zeeman splitting of $\Delta_{\mathrm{eff}} \approx 22$ μeV at the Dirac point (see Methods). As the TE–TM splitting is considerably larger than the Zeeman splitting in the lattice studied here, the experimentally determined bandgap that opens because of the magnetic field is $\Delta_{\mathrm{g}} = 108 \pm 32$ μeV with $\Delta_{\mathrm{eff}} < \Delta_{\mathrm{g}} < \beta_{\mathrm{eff}}$, which, as we show later, is

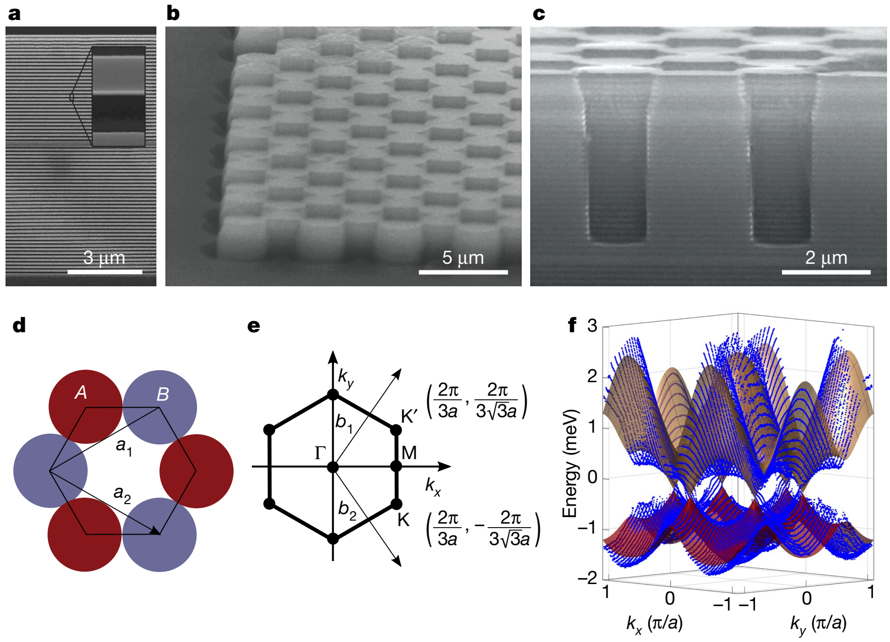


Fig. 2 | Lattice device layout and geometry. a–**c**, Scanning electron microscope images of the processed polariton honeycomb lattice. **a**, Cleaved cross-section of the microcavity before processing. **b**, Tilted view of the half-etched honeycomb lattice for pillars with a diameter of $d = 2.0\,\mu$m and an overlap $v = a/d = 0.85$. **c**, Cleaved cross-section after etching, showing that only the top DBR has been etched. **d**, Real-space honeycomb unit cell with two-element basis. **e**, First Brillouin zone of the honeycomb lattice, featuring the two non-equivalent K and K′ points. **f**, Measured Fourier-space energy-resolved photoluminescence of the investigated lattice. The blue data points are fitted to the measured dispersions, agreeing well with a tight-binding model (red and brown), accurately revealing the six Dirac cones at the K and K′ points.

reasonable for observing the topological features of our lattice and compares well with a gap size of about 100 µeV found for realistic system parameters[17].

After having established that a bandgap opens at the Dirac points under the influence of an external magnetic field, we use non-resonant excitation, under conditions allowing for polariton condensation into a chiral edge state within this bandgap. Scans of the photoluminescence signal in real space as well as Fourier space are performed on the polariton lattice at external magnetic fields of $B = -5$ T, 0 T and +5 T. The sample is excited on the zigzag edge by a pulsed and chopped laser beam with a diameter of 40 µm at 792 nm wavelength, tuned to the first stopband minimum (see Fig. 1a). In Fig. 3a and b, the linear polariton dispersion in the K′ → Γ → K direction at $B = 0$ T and at $B = +5$ T is displayed. The Dirac-cone dispersion is clearly visible ($P \approx 0.1$ mW). At a threshold power of $P_{th} = 1.8$ mW, we observe a nonlinear increase of the output power as well as a sudden decrease in linewidth (see Methods for details), establishing that a polariton condensate has formed at the K and K′ points at around $k_y \approx \pm 0.77\pi/a$, as displayed in Fig. 3c. We now perform mode tomography, scanning the real-space (x, y) landscape and measuring the photoluminescence energy E_{PL}. Figure 3d shows the spectrum of a line perpendicular to the zigzag edge, taken at the position indicated by the solid white line in Fig. 3g. The dashed white line depicts the edge of the sample. Besides an S-band condensate throughout the excited structure at $E_S = 1.4674$ eV, we observe an edge mode: a region of high intensity residing only at the outermost row of lattice sites spatially and spectrally at an energy of 1.4678 eV (white ellipse in Fig. 3d)—the expected topological bandgap.

Figure 3e shows the intensity pattern integrated over the energetic range of the trivial S-band mode centred at 1.4674 eV. Clearly, the condensate is relatively homogeneous over a large fraction of the lattice. The white overlaid lattice geometry and the microscopy image insets illustrate the position of the edge of the sample. Now, by changing the energy of the mode tomography to an energy $E_{edge} = 1.4678$ eV, within the topological gap (under a magnetic field of +5 T), the existence of edge states becomes unequivocal (at $x \approx 8$ µm in Fig. 3g). The photoluminescence at this energy originates predominantly from the outermost row of lattice sites, with almost no emission detected from the bulk of the lattice. The mode is in excellent agreement with the Bloch-mode calculations in Fig. 1e–g. In addition, the theoretical description within a Ginzburg–Landau-based model confirms that polariton condensation into the edge mode occurs (see Methods for details).

We now move on to study the robustness of the topological edge state. We do so by installing an artificial defect into the lattice at $y = 15$ µm (white circle in Fig. 3e–g). The defect is formed by leaving one of the sites on the zigzag edge of the honeycomb lattice unoccupied (see Methods). Normally, such a strong defect would cause scattering into the bulk, but here (see Fig. 3f, g) such scattering is suppressed, indicating that the transport of the topological edge state is immune to such defects. In addition, we perform mode tomography using non-resonant excitation with a large spot at the corner position of the sample. When plotting the energy of the topological edge mode $E_{edge} = 1.4678$ eV in Fig. 3f, one clearly observes that the mode extends around the corner from the zigzag into the armchair configuration, without any sign of backscattering or bulk scattering. The measurements at $B = -5$ T show very similar behaviour, but the transport is in the opposite direction. When the magnetic field is absent ($B = 0$ T), the edge mode vanishes completely (see Methods for details). The observation of the edge mode around the corner and especially its existence at the armchair edge, where topologically trivial honeycomb lattices do not have edge modes[9,27], prove that the edge mode we observe is indeed topological and is endowed with topological protection.

For further insight into the nature of these edge states, we analyse hyperspectral images ((k_x, k_y) versus E_{PL}) to identify the dominant propagation direction, with and without magnetic field. The results are displayed in Fig. 4. While in real space the modes can be clearly separated in energy, the integration over Fourier space results in the topological edge mode and the trivial bulk modes overlying one another. Figure 4a, b depicts the integrated intensities of the full S-band condensate (1.467–1.468 eV), including the energies associated with the edge state, for experiments at $B = +5$ T and -5 T, respectively. To analyse the directionality of polariton transport, the maximum peak intensities at the two maxima at $k_x \approx 0$ and $k_y \approx \pm 0.77\pi/a$ are extracted by identifying the central pixel of the peak and averaging the intensity

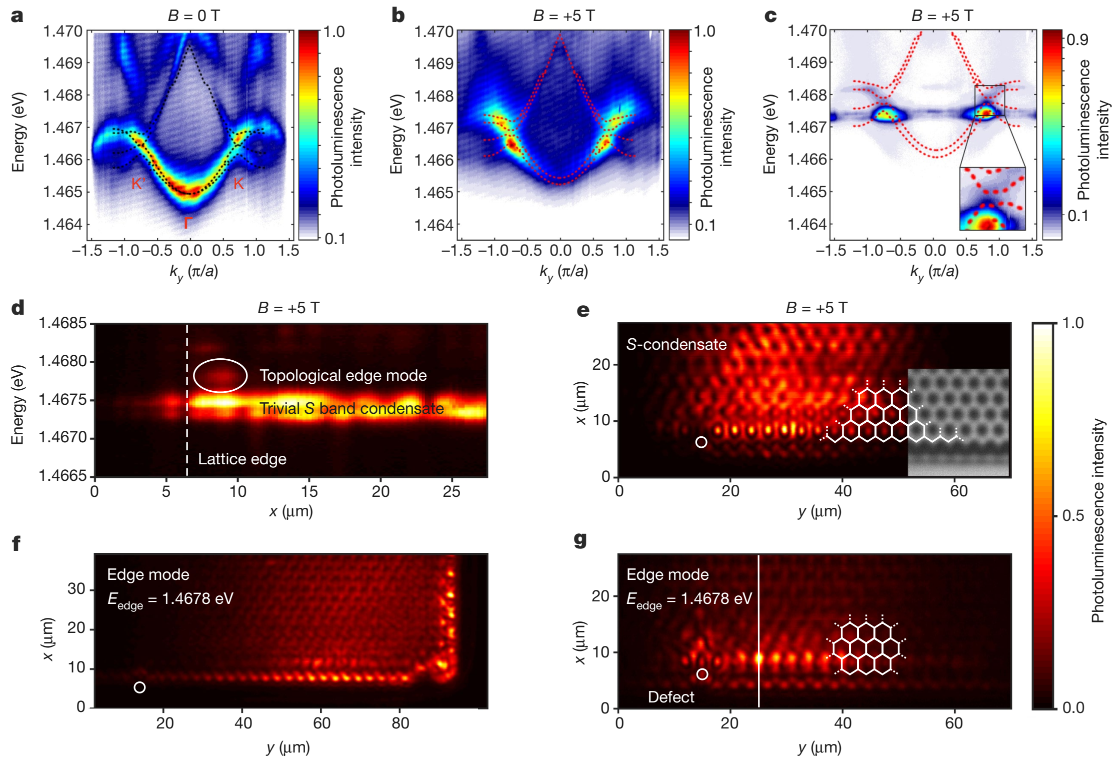


Fig. 3 | Photoluminescence measurements of a polariton condensate in a topological edge mode. a, Polariton dispersion along the K′ → Γ → K direction at $B = 0$ T compared with the calculated Bloch-mode model (black dots). **b**, **c**, Equivalent dispersions at $B = +5$ T, below (**b**) and above (**c**) the threshold $P_{th} = 1.8$ mW, where condensation into the K and K′ points of the *S* band is observed. Bloch-mode calculations (red dots) show a distinct gap. **d**, Real-space spectrum in the *x* direction along the straight white line in **g**. The *x* axis in **d**, **e** and **g** is the same. A trivial *S*-band condensate throughout the structure and a mode ($E = 1.4678$ eV) well separated from the bulk, located at the zigzag edge (dashed white line), are observed. **e**, Mode tomography displaying a homogeneous trivial *S*-band condensate ($E_S = 1.4673$–1.4675 eV) within the pump spot diameter of 40 μm. The inset shows a microscopy image of the structure. **f**, **g**, Mode tomography of the topological edge mode ($E_{edge} = 1.4678$ eV) at the corner position of the sample (**f**) and at the same position as in **e** (**g**). The mode is well located at the zigzag edge and clearly extends around the corner to the armchair configuration.

of a region of 3 × 3 pixels centred around this position. The vertical axis of Fig. 4c shows the ratio of the luminescence travelling in one direction ($+k_y$) to that in the opposite direction ($-k_y$). Deviation of this quantity from unity is an essential characteristic of a chiral state, and an opposite deviation should appear for opposite applied magnetic field. The corresponding intensity ratios are plotted in blue and show a clear directionality when an external magnetic field is applied. The transport changes its predominant direction along the edge when the direction of the magnetic field is inverted. This observation supports the interpretation of the edge-states being a result of a topologically non-trivial bandgap, with the edge mode contributing to the chirality along the edge. Reversing the magnetic field reverses the slope of the dispersion

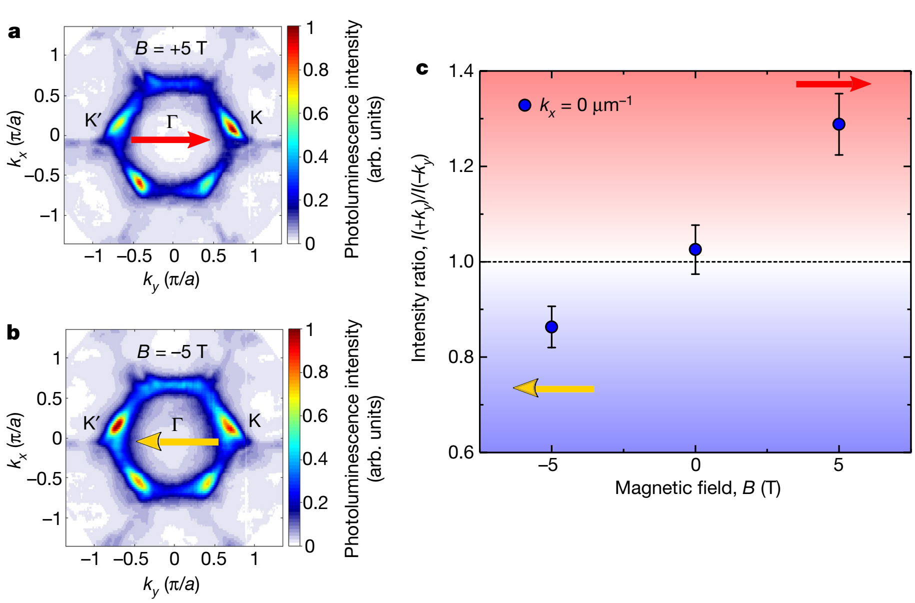


Fig. 4 | Chirality and propagation of the condensate. a, **b**, Spectroscopic hyperspectral measurement of the full *S*-band condensate (1.467–1.468 eV) at the K and K′ points, including the energies linked to the topological edge modes for $B = +5$ T (**a**) and $B = -5$ T (**b**). The zigzag edge is oriented in the *y* direction. **c**, Polariton intensity ratio between the K′ and K points in the k_y direction ($k_x \approx 0$) as a function of the applied magnetic field. The dominant propagation direction is inverted (yellow and red arrows) when the direction of the magnetic field is reversed. For $B = 0$ T, no dominant propagation direction is observed. The error bars originate from image distortions, inhomogeneities of the excitation and uncertainties during data processing, and are estimated at 5%.

curve of the topological edge mode, which is physically manifested in reversing the group velocity. On the other hand, we find no systematic directionality for the peaks at $k_x \approx -0.77\pi/a$ in Fig. 4a, b, which implies that these arise solely from the bulk condensate.

The experimental results depicted in Figs. 3, 4 unequivocally prove the observation of an exciton-polariton topological Chern insulator. The application of a magnetic field to the honeycomb lattice opens up a topological bandgap, with topological edge states supporting unidirectional transport whose propagation direction is determined by the field polarity. The lack of scattering from an artificial defect manifests the robustness of the topological edge mode. Furthermore, the observation of the edge mode extending around the corner and at the armchair termination without bulk scattering is a distinct feature of the topological edge mode, highlighting its topological protection against defects and disorder. Our results lead the way to efficient light trapping and topologically protected propagation of coherent exciton-polariton condensates in a well-developed semiconductor platform in which electrical driving can be envisaged[30]. We now aim to explore the topological lasing aspect of these experiments further, by comparing topological edge-mode lasing with lasing from a trivial edge mode in, for example, a Semenoff insulator. Such experiments would also link our exciton-polariton platform to the recently observed topological insulator laser[31,32]. Because of the interacting nature of polaritons, in-depth study of collective bosonic effects in topological insulators can be envisaged. For example, the large nonlinearity displayed by this exciton-polariton topological system can support the observation of solitons in topological insulators, which have been proposed[33,34] but not yet observed in any system. This work is a step towards new topological polaritonic devices with properties and functionalities involving nonlinearity, gain, interactions and coherence.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0601-5.

Received: 8 March 2018; Accepted: 13 August 2018;
Published online 8 October 2018.

1. Hasan, M. Z. & Kane, C. L. Colloquium: topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
2. Klitzing, K. v., Dorda, G. & Pepper, M. New method for high-accuracy determination of the fine-structure constant based on quantized Hall resistance. *Phys. Rev. Lett.* **45**, 494–497 (1980).
3. Haldane, F. D. M. Model for a quantum Hall effect without Landau levels: condensed-matter realization of the "parity anomaly". *Phys. Rev. Lett.* **61**, 2015–2018 (1988).
4. Kane, C. L. & Mele, E. J. Quantum spin Hall effect in graphene. *Phys. Rev. Lett.* **95**, 226801 (2005).
5. Bernevig, B. A., Hughes, T. L. & Zhang, S.-C. Quantum spin Hall effect and topological phase transition in HgTe quantum wells. *Science* **314**, 1757–1761 (2006).
6. König, M. et al. Quantum spin Hall insulator state in HgTe quantum wells. *Science* **318**, 766–770 (2007).
7. Haldane, F. D. M. & Raghu, S. Possible realization of directional optical waveguides in photonic crystals with broken time-reversal symmetry. *Phys. Rev. Lett.* **100**, 013904 (2008).
8. Wang, Z., Chong, Y. D., Joannopoulos, J. D. & Soljacić, M. Observation of unidirectional backscattering-immune topological electromagnetic states. *Nature* **461**, 772–775 (2009).
9. Rechtsman, M. C. et al. Photonic Floquet topological insulators. *Nature* **496**, 196–200 (2013).
10. Hafezi, M., Mittal, S., Fan, J., Migdall, A. & Taylor, J. M. Imaging topological edge states in silicon photonics. *Nat. Photon.* **7**, 1001–1005 (2013).
11. Jotzu, G. et al. Experimental realization of the topological Haldane model with ultracold fermions. *Nature* **515**, 237–240 (2014).
12. Aidelsburger, M. et al. Measuring the Chern number of Hofstadter bands with ultracold bosonic atoms. *Nat. Phys.* **11**, 162–166 (2015).
13. Yang, Z. et al. Topological acoustics. *Phys. Rev. Lett.* **114**, 114301 (2015).
14. Fleury, R., Khanikaev, A. B. & Alù, A. Floquet topological insulators for sound. *Nat. Commun.* **7**, 11744 (2016).
15. Süsstrunk, R. & Huber, S. D. Observation of phononic helical edge states in a mechanical topological insulator. *Science* **349**, 47–50 (2015).
16. Karzig, T., Bardyn, C.-E., Lindner, N. H. & Refael, G. Topological polaritons. *Phys. Rev. X* **5**, 031001 (2015).
17. Bardyn, C.-E., Karzig, T., Refael, G. & Liew, T. C. H. Topological polaritons and excitons in garden-variety systems. *Phys. Rev. B* **91**, 161413(R) (2015).
18. Nalitov, A. V., Solnyshkov, D. D. & Malpuech, G. Polariton $\mathbb{Z}$ topological insulator. *Phys. Rev. Lett.* **114**, 116401 (2015).
19. Weisbuch, C., Nishioka, M., Ishikawa, A. & Arakawa, Y. Observation of the coupled exciton-photon mode splitting in a semiconductor quantum microcavity. *Phys. Rev. Lett.* **69**, 3314–3317 (1992).
20. Kasprzak, J. et al. Bose–Einstein condensation of exciton polaritons. *Nature* **443**, 409–414 (2006).
21. Carusotto, I. & Ciuti, C. Quantum fluids of light. *Rev. Mod. Phys.* **85**, 299–366 (2013).
22. Schneider, C. et al. Exciton-polariton trapping and potential landscape engineering. *Rep. Prog. Phys.* **80**, 016503 (2017).
23. St-Jean, P. et al. Lasing in topological edge states of a one-dimensional lattice. *Nat. Photon.* **11**, 651–656 (2017).
24. Peleg, O. et al. Conical diffraction and gap solitons in honeycomb photonic lattices. *Phys. Rev. Lett.* **98**, 103901 (2007).
25. Jacqmin, T. et al. Direct observation of Dirac cones and a flatband in a honeycomb lattice for polaritons. *Phys. Rev. Lett.* **112**, 116402 (2014).
26. Milićević, M. et al. Edge states in polariton honeycomb lattices. *2D Mater.* **2**, 034012 (2015).
27. Fujita, M., Wakabayashi, K., Nakada, K. & Kusakabe, K. Peculiar localized state at zigzag graphite edge. *J. Phys. Soc. Jpn* **65**, 1920–1923 (1996).
28. Castro Neto, A. H., Guinea, F., Peres, N. M. R., Novoselov, K. S. & Geim, A. K. The electronic properties of graphene. *Rev. Mod. Phys.* **81**, 109–162 (2009).
29. Yao, Y., Ye, F., Qi, X.-L., Zhang, S.-C., and Fang, Z. Spin–orbit gap of graphene: first-principles calculations. *Phys. Rev. B* **75**, 041401(R) (2007).
30. Suchomel, H. et al. A plug and play platform for electrically pumped polariton simulators and topological lasers. Preprint at https://arxiv.org/abs/1803.08306 (2018).
31. Bahari, B. et al. Nonreciprocal lasing in topological cavities of arbitrary geometries. *Science* **358**, 636–640 (2017).
32. Bandres, M. A. et al. Topological insulator laser part II: experiments. *Science* **359**, aar4005 (2018).
33. Lumer, Y., Plotnik, Y., Rechtsman, M. C. & Segev, M. Self-localized states in photonic topological insulators. *Phys. Rev. Lett.* **111**, 243905 (2013).
34. Kartashov, Y. V. & Skryabin, D. V. Modulational instability and solitary waves in polariton topological insulators. *Optica* **3**, 1228–1236 (2016).

Acknowledgements We thank R. Thomale for discussions. S.K. acknowledges the European Commission for the H2020 Marie Skłodowska-Curie Actions (MSCA) fellowship (Topopolis). S.K., S.H. and M.S. are grateful for financial support by the JMU-Technion seed money programme. S.H. also acknowledges support by the EPSRC "Hybrid Polaritonics" grant (EP/M025330/1). The Würzburg group acknowledges support by the ImPACT Program, Japan Science and Technology Agency and the ENB programme (Tols 836315) of the State of Bavaria. T.C.H.L. and R.G. were supported by the Ministry of Education (Singapore) grant no. 2017-T2-1-001.

Author contributions S.K., M.S., C.S. and S.H. initiated the study and guided the work. S.K., T.H., K.W., M.E. and S.H. designed and fabricated the device. S.K. and T.H. performed optical measurements. S.K., T.H., O.A.E. and C.S. analysed and interpreted the experimental data. O.A.E., R.G., T.C.H.L., M.A.B. and M.S. developed the theory. S.K., T.H., O.A.E., T.C.H.L., C.S., M.S. and S.H. wrote the manuscript, with input from all co-authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0601-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.K. or S.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Basic sample characterization. Honeycomb lattices of coupled micropillars were etched into a planar semiconductor Fabry–Pérot microcavity, grown by molecular beam epitaxy. The cavity consists of a GaAs λ-cavity equipped with three 16-nm-wide $In_{0.04}Ga_{0.96}As$ quantum wells sandwiched between two distributed Bragg reflectors with 30 (35.5) $Al_{0.10}Ga_{0.90}As$/AlAs top (bottom) mirror pairs. The quantum wells were placed at the maximum of the electromagnetic field, resulting in a strong exciton-photon coupling with a Rabi splitting of 4.3 meV (see Extended Data Fig. 1). The set of quantum wells emit with an exciton linewidth of $\gamma_X = 1.2$ meV (full width at half maximum), measured with the top mirror etched away. The quality factor of the cavity was determined experimentally by measuring the mode linewidth of a single large pillar at a negative (photonic) detuning, $\delta \approx -11.5$ meV, to be about 13.000 ($\gamma_C = 0.11$ meV). The overall layer thicknesses decrease radially towards the outside of the wafer, affecting mainly the photonic mode. The decrease in thickness allows the experimenter to choose a certain exciton-to-photon detuning. By increasing the asymmetry between the cavity mode and the stopband centre, the TE–TM splitting of the cavity mode is increased, mimicking a spin–orbit interaction[35]. However, the further the cavity mode is moved towards the edge of the stopband, the lower the reflectivity becomes, implying a lower Q-factor and larger polariton linewidth. Based on simulations and experience from previously grown samples, a DBR (cavity) asymmetry factor of 1.03 (0.91) was chosen for the microcavities presented in this work (1.00 (1.00) would be the symmetric case), yielding a TE–TM splitting of approximately 600 μeV at $k_\parallel = 2.0\,\mu\text{m}^{-1}$ (see Extended Data Fig. 2). Both the indium content in the InGaAs quantum well and the quantum-well thickness were optimized with regards to the smallest linewidth and largest Zeeman splitting. 4% indium and 16 nm thickness yield a Zeeman splitting $\Delta E_Z = 540$ μeV at $B = 5$ T (see Extended Data Fig. 2). After sample growth, a honeycomb lattice of approximately $(120 \times 120)\,\mu\text{m}^2$ with a pillar diameter of $d = 2.0\,\mu$m and an overlap $v = a/d = 0.85$ was defined using an electron beam lithography process and subsequent wet etching. Extended Data Fig. 3 shows a microscope image of an intended defect in the form of a site missing in a zigzag edge of the honeycomb lattice (indicated by the red arrow).

Photoluminescence experiment. Non-resonant photoluminescence experiments with and without an applied magnetic field were carried out using a linearly polarized, pulsed titanium-sapphire laser with a repetition rate of 82 MHz and a pulse length of approximately 2 ps. The wavelength of the laser was set to be $\lambda_L = 792$ nm, coinciding with the first high-energy stopband minimum of the microcavity structure.

The emission was collected using a microscope objective (20×, NA = 0.4) and imaged at the entrance slit of a Czerny–Turner spectrometer, equipped with a charge-coupled device (CCD) camera with a resolution of about 20 μeV. A motorized imaging lens allows for automated hyperspectral images ((k_x, k_y) versus E) and mode tomographies ((x, y) versus E).

The sample was mounted in a liquid helium flow cryostat (Oxford Instruments Microstat), operating at a temperature of $T = 4$ K. Using superconducting coils, a magnetic field $B = -5$ T to $+5$ T can be applied in Faraday geometry.

Bloch-mode calculations for polariton honeycomb lattices. A widely accepted model describes the dynamics of excitons with spin-up (ψ^+) and spin-down (ψ^-) coupled to cavity photons carrying the right (E^+) and the left (E^-) circular polarizations, respectively[18,34,36], and is governed by

$$i\hbar\partial_t E^{\pm} = \left(-\frac{\hbar^2}{2m_C}\nabla_\perp^2 + V(\boldsymbol{r}) + \omega_C - i\gamma\right)E^{\pm} + \beta(\partial_x \mp i\partial_y)^2 E^{\mp} + \hbar\Omega_R\psi^{\pm} + E_p^{\pm}e^{-i\omega_p t} \quad (1)$$

$$i\hbar\partial_t\psi^{\pm} = \left(\omega_E - i\gamma \pm \frac{\Delta_B}{2}\right)\psi^{\pm} + \hbar\Omega_R E^{\pm} \quad (2)$$

(where ∂_t represents partial derivative with respect to time t, and E_p is the external coherent radiation explained below). Here, the normalization is such that $|E^{\pm}|^2$ and $|\psi^{\pm}|^2$ are the number of particles per unit area. The quantities ω_C and ω_E represent the energies of bare photons and excitons, respectively. In the present configuration, the photon–exciton detuning is negative, $\delta = (\omega_C - \omega_E) = -6$ meV. The photon–exciton coupling strength is given by the parameter $\hbar\Omega_R$, which defines the Rabi splitting as $2\hbar\Omega_R = 4.5$ meV. Here, $m_C = 32.3 \times 10^{-6} m_e$ is the effective photon mass in the planar region, and m_e is the free electron mass. The effective mass of excitons is $m_E \approx 10^5 m_C$. An external photonic potential $V(\boldsymbol{r})$ is defined within the unit cell of the honeycomb structure, constructed of circular mesas (micropillars). We assume that the potential is $V(\boldsymbol{r}) = 30$ meV outside the mesas and zero otherwise. In what follows, we assume that the intensity of polaritons is weak enough and thus neglect nonlinear interactions between them. The TE–TM splitting of the cavity modes gives rise to the linear coupling between right- and left- circular polarizations and is denoted by β (ref. [36]). We account for the magnetic field via Zeeman splitting Δ_B of the excitonic states in the quantum wells.

Equations (1) and (2) allow for accurate simulation of the propagation of the chiral modes coherently injected into the system by means of the external coherent radiation ($E_p^{\pm}$), with an appropriately chosen frequency (ω_p) within the topological gap. First, we calculate the energy–momentum band structure of the honeycomb lattices, using a full description of the Bloch modes that takes into account all relevant system parameters. For this aim, we solve the following eigenvalue problem for the energy $\hbar\mu(\boldsymbol{k}_b)$ of the Bloch mode with the Bloch vector $\boldsymbol{k}_b = (k_{bx}, k_{by})$

$$\hbar\mu(\boldsymbol{k}_b)\begin{pmatrix}\boldsymbol{p}_b^+(\boldsymbol{r},\boldsymbol{k}_b)\\ \boldsymbol{p}_b^-(\boldsymbol{r},\boldsymbol{k}_b)\end{pmatrix} = \begin{pmatrix}\hat{L}^+ & \hat{C}^+\\ \hat{C}^- & \hat{L}^-\end{pmatrix}\begin{pmatrix}\boldsymbol{p}_b^+(\boldsymbol{r},\boldsymbol{k}_b)\\ \boldsymbol{p}_b^-(\boldsymbol{r},\boldsymbol{k}_b)\end{pmatrix} \quad (3)$$

The circularly polarized polaritonic wavefunctions are $\boldsymbol{p}_b^{\pm}(\boldsymbol{r},\boldsymbol{k}_b) = (c_b^{\pm}(\boldsymbol{r},\boldsymbol{k}_b), x_b^{\pm}(\boldsymbol{r},\boldsymbol{k}_b))$, where the functions $c_b^{\pm}(\boldsymbol{r},\boldsymbol{k}_b)$ and $x_b^{\pm}(\boldsymbol{r},\boldsymbol{k}_b)$ describe the amplitude distributions of the photonic and excitonic component of the Bloch modes in real space, defined in the plane of the microcavity $\boldsymbol{r} = (x, y)$. The diagonal of the matrix in equation (3) describes the single-particle coupled states of excitons and photons and is given by the expression

$$\hat{L}^{\pm} = \begin{pmatrix}\omega_C + V(\boldsymbol{r}) - \frac{\hbar^2}{2m_C}(\boldsymbol{\nabla}_\perp + i\boldsymbol{k}_b)^2 & \hbar\Omega_R\\ \hbar\Omega_R & \omega_E - \frac{\hbar^2}{2m_E}(\boldsymbol{\nabla}_\perp + i\boldsymbol{k}_b)^2 \pm \frac{\Delta_B}{2}\end{pmatrix} \quad (4)$$

The coupling between both polarization components is given by the matrices $\hat{C}^{\pm}$

$$\hat{C}^{\pm} = \begin{pmatrix}\beta[(\boldsymbol{\nabla}_\perp + i\boldsymbol{k}_b)\cdot(\boldsymbol{e}_x \mp i\boldsymbol{e}_y)]^2 & 0\\ 0 & 0\end{pmatrix} \quad (5)$$

which includes the TE–TM splitting of the photonic modes. Δ_B describes the Zeeman splitting of the excitons due to the applied magnetic field. The unit vectors $\boldsymbol{e}_{x,y}$ shows the directions of Cartesian coordinates (x, y) in the plane of the microcavity. To reduce computational efforts greatly, it is convenient to solve the eigenvalue problem in polaritonic basis after diagonalization of the matrix $\hat{L}^{\pm}$, describing the coupling between photons and excitons. In this case, the effective parameter for the TE–TM splitting $\beta_{\text{eff}} \approx \beta|C|^2$ and the Zeeman splitting $\Delta_{\text{eff}} \approx \Delta_B|X|^2$ are scaled with the photonic and excitonic components of the polaritons, respectively (where $|C|^2$ and $|X|^2$ are the Hopfield coefficients, calculated for the respective Bloch mode).

Physically, the excitation occurs by coherent illumination with linearly polarized light at the frequency within the topological gap. The seeding pulse duration is about 40 ps. Other parameters: $\Delta_B = \pm 0.8$ meV, $\beta = 0.20$ meV μm², $\beta_{\text{eff}} = 0.15$ meV μm², $\gamma = 0.01$ meV, polariton lifetime $\tau = 35$ ps, pillar diameter $d = 2.0\,\mu$m, centre-to-centre separations $a = 2\,\mu$m.

Polariton chiral edge mode propagating around a corner and defect. Using the Bloch mode calculations described above, we now excite the system resonantly and calculate its evolution in time for a Zeeman splitting of $\Delta_B = +0.8$ meV. The polaritons propagate along the edge and around the 90° corner as expected for a topological edge mode (see Extended Data Fig. 4). In the same way, we find that the edge mode avoids an artificial defect in the form of a site missing in a zigzag chain (see Extended Data Fig. 5).

Tight-binding model. A tight-binding model describing the artificial graphene band structure is given by

$$E_{\text{hc}}(\boldsymbol{k}_\parallel) = E_0 \pm t\sqrt{3 + f(\boldsymbol{k}_\parallel)} - t' f(\boldsymbol{k}_\parallel) \quad (6)$$

$$f(\boldsymbol{k}_\parallel) = 2\cos(\sqrt{3}k_y a) + 4\cos\left(\frac{\sqrt{3}}{2}k_y a\right)\cos\left(\frac{3}{2}k_x a\right) \quad (7)$$

Here, t is the nearest neighbour, t' is the next-nearest neighbour and $\boldsymbol{k}_\parallel$ is the measured in-plane lattice vector, which is linked to the measured polariton emission angle θ by $k_\parallel = (\omega/c)\sin\theta$. The resulting band structure of this model is presented as a fit to the data in Fig. 2f, in red and yellow-brown.

Gap measurement. When an external magnetic field is applied to this lattice, a bandgap is predicted to open at the Dirac points. We make use of a $\lambda/4$ polarization series to detect this gap. In Extended Data Fig. 6, exemplary images of the $\lambda/4$-series at external magnetic fields of $B = 0$ T and 5 T are presented. The energetic position of the Dirac point was evaluated by fitting a Lorentzian peak profile to the line spectrum through the Dirac point at K′. The peak positions were plotted against the angle of the $\lambda/4$-waveplate. Assuming that the position of

the peak corresponds linearly to its polarization, the resulting graph can be fitted using the equation

$$I(\phi)=\frac{1}{2}[S_0+S_1\cos^2(2\phi)+S_2\sin(2\phi)\cos(2\phi)+S_3\sin(2\phi)]$$

normally used to fit the peak intensity in a $\lambda/4$-series with a constant peak position. At $B=+5$ T, a bandgap of $E_g=108\pm32\,\mu$eV was evaluated. The peak movement with no external magnetic field applied can be attributed to imperfections of the $\lambda/4$-waveplate. It was used to indicate the uncertainty of the bandgap that opens when an external magnetic field is applied.

Input/output characteristics and linewidth. By increasing the power of the non-resonant pump laser, measuring the intensity and linewidth we determine the threshold characteristic of the polaritons as shown in Extended Data Fig. 7. At a typical threshold of $P_{th}\approx1.8$ mW a distinct nonlinear increase in intensity as well as a sudden decrease in linewidth can be observed. The mode associated with the gap mode becomes visible around $P=P_{th}$.

Ginzburg–Landau calculations of polariton condensation into a topological edge mode. In theoretical proposals such as in refs [16–18], no particular form of excitation was considered explicitly. It is clear from the theoretical Bloch mode calculations that a resonant excitation is able to directly populate the various modes. For the case of non-resonant excitation, further theoretical work is needed to access the physics of polariton condensation into a chiral topological edge mode, described in Fig. 3.

Here we use the driven-dissipative Ginzburg–Landau model, frequently used to describe the spatial form of polariton condensates:

$$i\hbar\frac{\partial}{\partial t}\begin{pmatrix}\psi_+\\ \psi_-\end{pmatrix}=\begin{pmatrix}\hat{L}_0+\frac{\Delta_{\text{eff}}}{2}-i|\psi_+|^2 & \beta_{\text{eff}}(\hat{k}_x-i\hat{k}_y)^2\\ \beta_{\text{eff}}(\hat{k}_x+i\hat{k}_y)^2 & \hat{L}_0-\frac{\Delta_{\text{eff}}}{2}-i|\psi_-|^2\end{pmatrix}\begin{pmatrix}\psi_+\\ \psi_-\end{pmatrix}\quad(8)$$

Here ψ_+ and ψ_- denote the two spin components of the two-dimensional polariton wavefunction. The operator $\hat{L}_0$ is defined as $\hat{L}_0=\hat{E}_0+V(\boldsymbol{r})+iW(\boldsymbol{r})$, where the operator $\hat{E}_0$ corresponds to the bare kinetic energy of polaritons, modelled with an effective mass m. The potential $V(\boldsymbol{r})$ accounts for the honeycomb lattice structure and Δ_{eff} is the Zeeman splitting caused by the applied magnetic field. The two spin components are coupled by the off-diagonal spin–orbit coupling term, physically corresponding to a TE–TM splitting with strength β_{eff}.

The linear gain and loss in the system is described by the term $W(\boldsymbol{r})=P(\boldsymbol{r})-\Gamma(\boldsymbol{r})$, where $P(\boldsymbol{r})$ represents the spatially dependent non-resonant pumping and $\Gamma(\boldsymbol{r})$ is a spatially dependent dissipation rate. We assume that the loss is higher outside the micropillar regions. For the system to form a steady state, it is important to consider also nonlinear loss terms. For simplicity, we neglect the effect of polariton–polariton interactions, which is valid provided we operate not too far above the condensation threshold.

Before solving the full nonlinear problem, it is instructive to consider the spectrum of linearized modes, obtained by neglecting the nonlinear term. Considering a strip geometry, where the potential is periodic in the x direction and bounded in the orthogonal y direction, the Bloch theory can be used to obtain the complex band structure of the system eigenmodes with respect to a wavevector k_x. The result is shown in Extended Data Fig. 8a. For parameters comparable with the experiment, we obtain a topological bandgap bridged by chiral edge states. A slight asymmetry in the edge states occurs owing to the pumping $P(\boldsymbol{r})$, which is taken to be periodic in the x direction (to maintain validity of the Bloch theory) but preferentially localized near the bottom edge of the strip.

Extended Data Fig. 8b shows the imaginary parts of the same eigenmodes. The chiral edge state is found to have the largest imaginary component, meaning that it has a larger gain than other states and would be preferentially selected during polariton condensation. However, it should be noted that in principle there are an infinite number of energy states obtainable with the present theory, and it is necessary in practice to introduce an energy cut-off to solve the Bloch Hamiltonian.

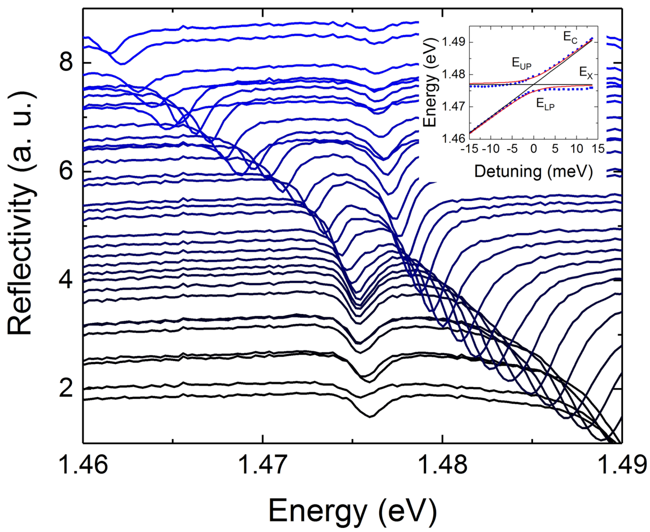
To rule out the potential population of higher energy states, we rely on solution of the full nonlinear problem, propagating equation (8) in time to a steady state. Here we do indeed find condensation in a chiral edge state, as illustrated with the colour scale in Extended Data Fig. 8a. A slight shift from the results of the linear band-structure calculation occurs because of the nonlinear term. Extended Data Fig. 8c shows the obtained wavefunction in real space, which is indeed localized at the edge excited by the pumping, agreeing with the experimental findings. In addition to explaining our experimental observation of chiral current under non-resonant pumping, these theoretical results predict a polariton lasing in a topological edge state: a topological polariton laser[31,32]. However, we point out that a strict comparison between a topological laser and a non-topological laser (as in ref. [32]) cannot easily be made, as the topological gap cannot be closed without changing a wide range of other system parameters.

Real-space mode tomographies at $B=0$ T and $B=-5$ T. For the sake of completeness, we perform the same mode tomographies at $B=0$ T and $B=-5$ T as were displayed in Fig. 3 for $B=+5$ T. The results are displayed in Extended Data Fig. 9a–d ($B=0$ T) and Extended Data Fig. 9e–h ($B=-5$ T). Whereas for the $B=0$ T case no distinct edge mode is observed, the behaviour for $B=-5$ T is qualitatively similar to that for $B=+5$ T, as expected.

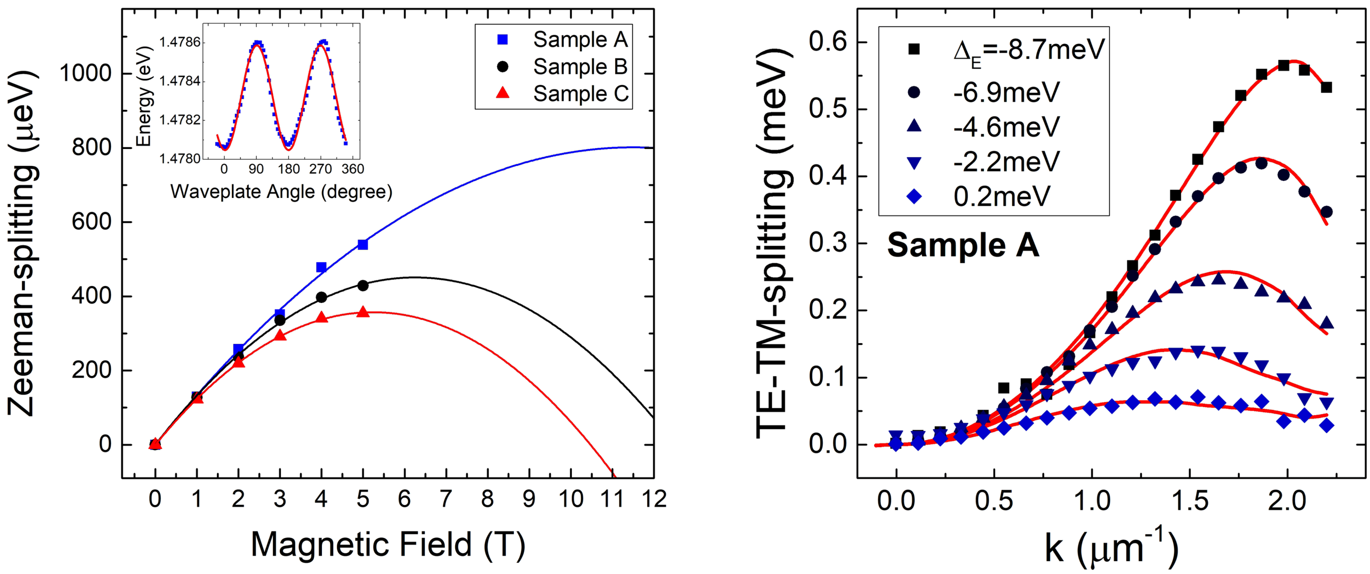
Data availability

The research data that support this publication can be accessed at https://doi.org/10.17630/4a62cbdd-bcae-45d7-a556-3cda53c0a656. Additional data related to this paper may be requested from the corresponding authors.

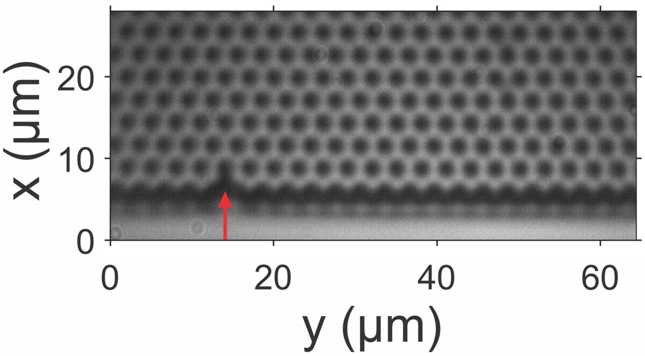
35. Sala, V. G. et al. Spin–orbit coupling for photons and polaritons in microstructures. *Phys. Rev. X* **5**, 011034 (2015).
36. Deveaud, B. *The Physics of Semiconductor Microcavities* (Wiley-VCH, Weinheim, 2007).



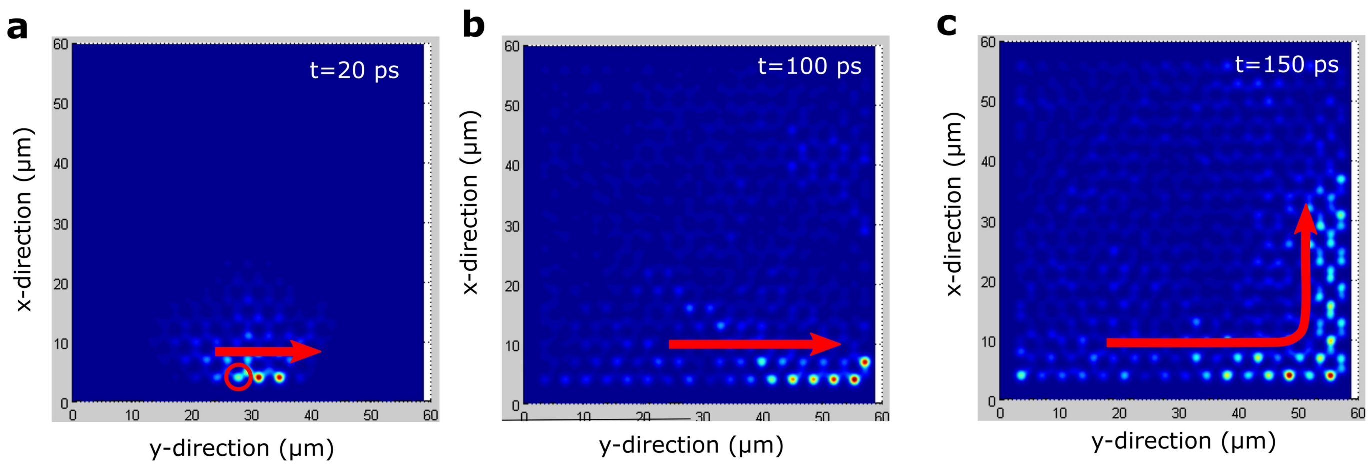
Extended Data Fig. 1 | White-light reflectivity measurements as a function of the detuning. Reflectivity measurements are shown as a function of the detuning. A distinct anticrossing behaviour with a Rabi splitting of $2\hbar\Omega_R = 4.3$ meV can be observed. The measurements were performed on a sample piece with approximately 15 mirror pairs removed from the top DBR to increase the signal quality. Inset, fitted peak positions versus detuning. UP and LP stand for upper polariton and lower polariton.



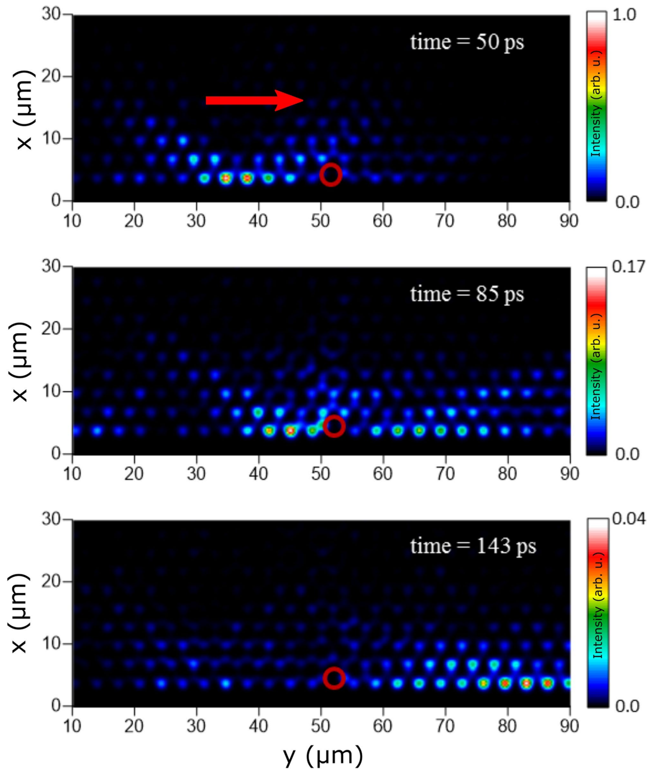
Extended Data Fig. 2 | Zeeman splitting and TE–TM splitting for a III–V microcavity hosting $In_{0.04}Ga_{0.96}As$ quantum wells. Left, Zeeman splitting with regard to the magnetic field, including second-order polynomial fits as a guide to the eye. Sample A is the one used in this experiment. Inset, example of central emission energies for a $\lambda/4$-series with a sine fit. Right, experimentally determined TE–TM splitting at various detunings Δ_E for sample A including fits with modified photonic Hopfield coefficients (red).



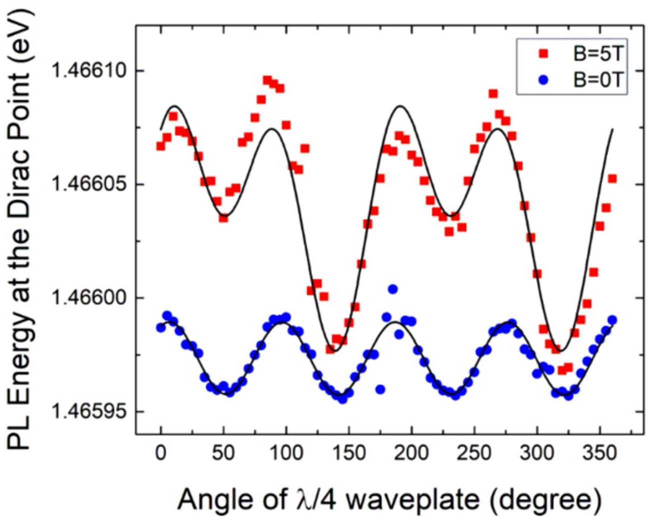
Extended Data Fig. 3 | Microscopy image of a zigzag-edge polariton honeycomb lattice with an intentional defect. Shown is an image of the honeycomb lattice with $d = 2.0$ µm and $v = 0.85$, with an intentional defect in the zigzag chain, selected for experiments in an external magnetic field.



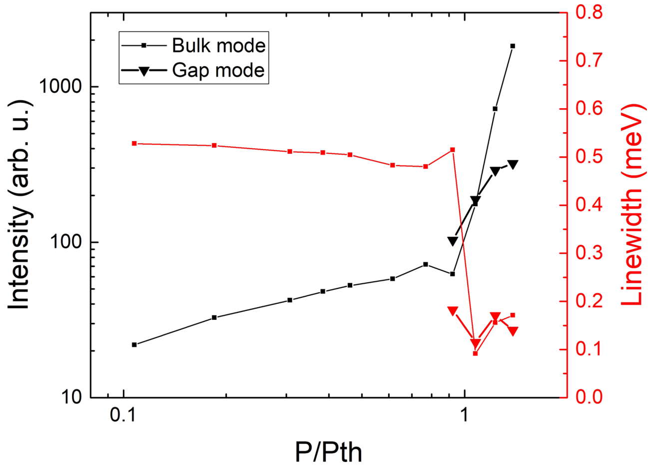
Extended Data Fig. 4 | Polariton chiral edge mode propagating around a corner. a–c, Propagation dynamics of edge modes injected coherently into the topological gap and calculated within model 1. Shown here is the right-moving propagation for the positive-value splitting $\Delta_B = +0.8$ meV ($|\Delta_{eff}| = 0.2$ meV) and $\beta = 0.20$ meV μm^2 ($\beta_{eff} = 0.15$ meV μm^2). A linearly polarized narrow coherent seeding beam injects both polarization components into the region marked by the red circle. At $t \approx 100$ ps, the mode propagates around the corner from the zigzag edge into the armchair one.



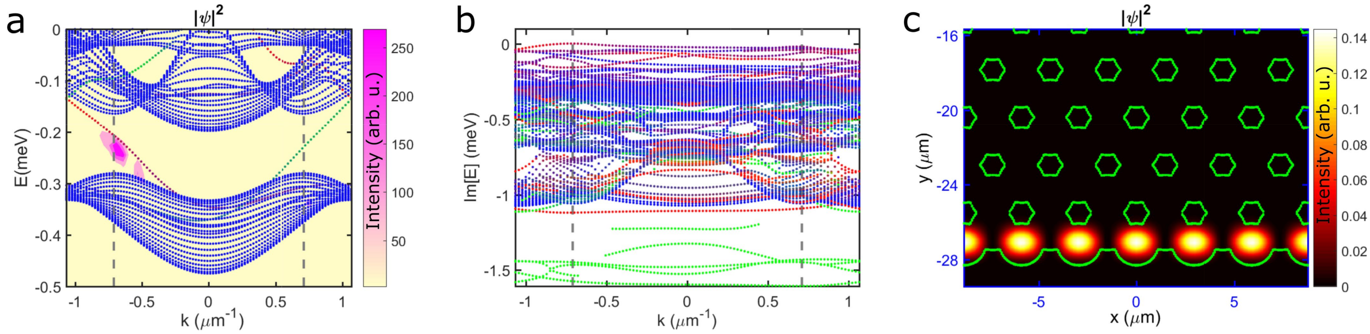
Extended Data Fig. 5 | Polariton chiral edge mode propagating and avoiding a defect. Propagation dynamics of edge modes injected coherently into the topological gap and calculated within model 1 are shown for the right-moving propagation for the positive-value splitting $\Delta_B = +0.8$ meV ($|\Delta_{eff}| = 0.2$ meV) and $\beta = 0.20$ meV μm^2 ($\beta_{eff} = 0.15$ meV μm^2). A linearly polarized narrow coherent seeding beam injects both polarization components into the mesa. At $t \approx 85$ ps, the mode propagates around the defect in the zigzag chain, marked by the red circle.



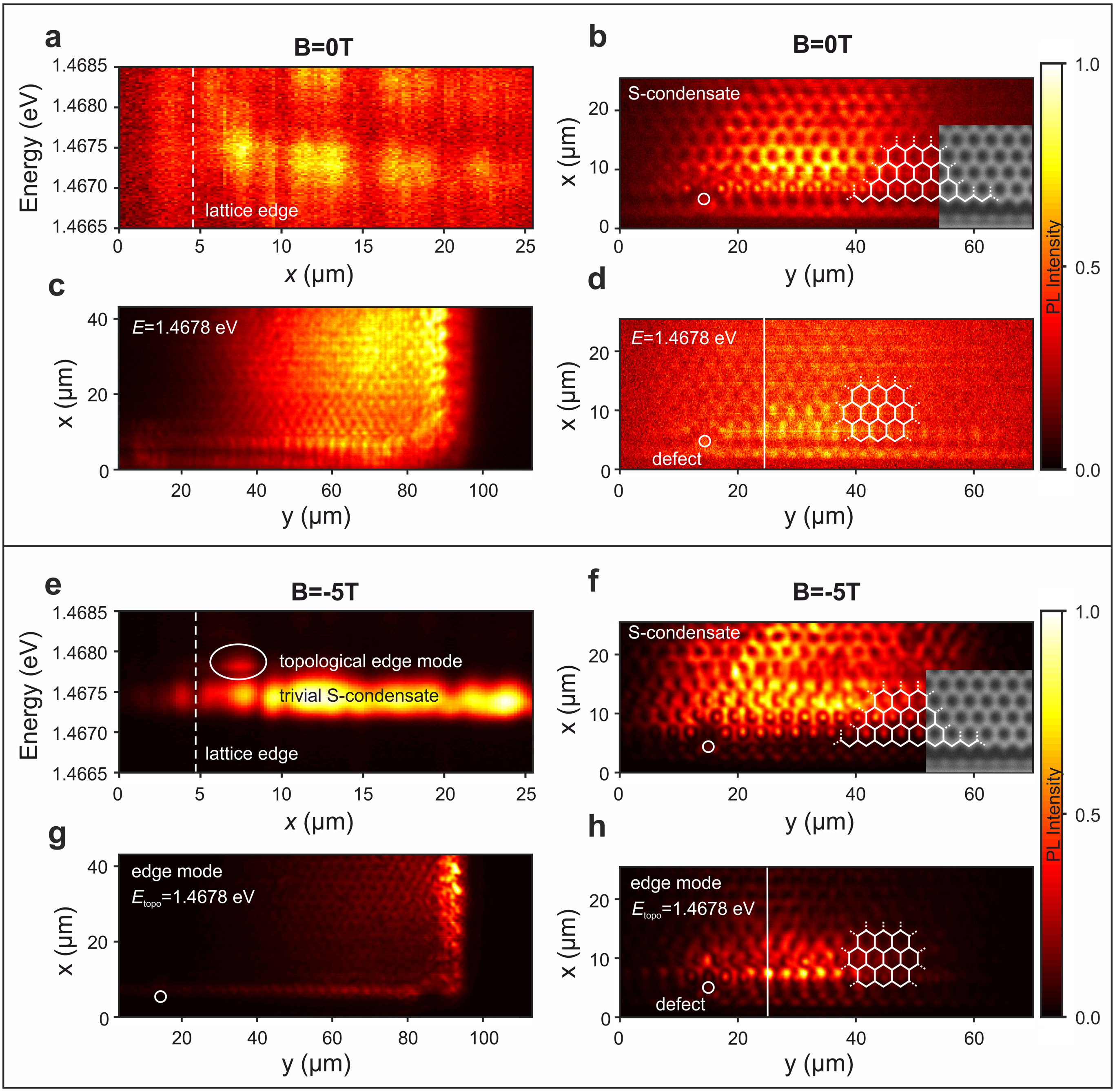
Extended Data Fig. 6 | Topological gap measurement. A λ/4-plate measurement at $B = 0$ T (blue) and $B = +5$ T (red) at the K point yields a bandgap of $E_g = 108 \pm 32\,\mu$eV. PL, photoluminescence.



Extended Data Fig. 7 | Input/output characteristics and linewidth behaviour as a function of pump power. Below threshold, the gap and bulk mode cannot be distinguished. At a typical threshold $P_{th} \approx 1.8$ mW, a distinct nonlinear increase in intensity as well as a sudden decrease in linewidth can be observed. Here, the populated gap modes show similar behaviour to the bulk mode.



Extended Data Fig. 8 | Driven-dissipative Gross–Pitaevskii calculation of polariton condensation into topological edge mode. a, Band structure of polaritons in a honeycomb lattice. The dotted curves represent the dispersion of the linear eigenmodes of a strip, colour-coded to represent localization on the bottom edge (red), upper edge (green) and in the bulk (blue). The shaded region represents the energy and momentum of the polariton steady state obtained from solving the driven-dissipative Gross–Pitaevskii equation. **b**, Imaginary components of the linear eigenmodes. The largest imaginary part corresponds to an edge state (the colour coding is the same as in **a**), suggesting that the edge state is most likely to be populated with increasing pumping. **c**, Edge state obtained from solution of the driven-dissipative Gross–Pitaevskii equation. Parameters: $\Delta_{\mathrm{eff}} = 0.3$ meV, $\beta_{\mathrm{eff}} = 0.2$ meV μm². The effective mass m was taken as 1.3×10^{-4} of the free electron mass; the potential of depth 0.5 meV was constructed from a honeycomb lattice of cylinders of radius 1 μm and centre-to-centre separation 1.7 μm; the pump spot was taken as a Gaussian centred on the strip edge with extent 7.5 μm in the y direction. A spatially uniform decay rate of 0.2 meV was supplemented with a 1.7 meV decay in the region outside the cylinders.



Extended Data Fig. 9 | Real-space mode tomographies of a polariton condensate at $B=0$ T and $B=-5$ T. a–d, Measurements at $B=0$ T. **a**, Real-space spectrum in x direction perpendicular to the zigzag edge along the straight white line in **d**. The real-space x axis is consistent between **a**, **b** and **d**. The dashed white line marks the physical edge of the lattice. Only a trivial S-band condensate can be observed throughout the structure. **b**, Mode tomography displaying the topologically trivial S-band condensate at $E_S = 1.4673$–1.4675 eV. A relatively homogeneous condensate within the pump spot diameter of 40 μm is observed. The inset shows a microscopy image of the structure. **c**, **d**, Mode tomography of the energy $E_{edge} = 1.4678$ eV for comparison at the corner position (**c**) and at the edge (**d**) of the sample. Without magnetic field, no localized edge mode can be observed. **e–h**, Measurements at $B=-5$ T (fully analogous to Fig. 3d–g). **e**, Real-space spectrum in the x direction perpendicular to the zigzag edge along the straight white line in **h**. The real-space x axis is consistent between **e**, **f** and **h**. The dashed white line marks the physical edge of the lattice. A trivial S-band condensate can be observed throughout the structure. At $E=1.4678$ eV, again we observe the appearance of a localized mode, well separated from the bulk and located at the zigzag edge. **f**, Mode tomography displaying the topologically trivial S-band condensate at $E_S = 1.4673$–1.4675 eV. A relatively homogeneous condensate within the pump spot diameter of 40 μm is observed. The inset shows a microscopy image of the structure. **g**, **h**, Mode tomography of the topological edge mode at $E_{edge} = 1.4678$ eV at the corner position (**g**) and at the edge (**h**) of the sample, showing clearly that the mode extends around the corner from the zigzag to the armchair configuration and avoids the intentional defect, both without bulk scattering.

LETTER

https://doi.org/10.1038/s41586-018-0618-9

In-plane anisotropic and ultra-low-loss polaritons in a natural van der Waals crystal

Weiliang Ma[1,11], Pablo Alonso-González[2,11]*, Shaojuan Li[1,11], Alexey Y. Nikitin[3,4], Jian Yuan[1], Javier Martín-Sánchez[2], Javier Taboada-Gutiérrez[2], Iban Amenabar[5], Peining Li[5], Saül Vélez[5,6], Christopher Tollan[5], Zhigao Dai[7], Yupeng Zhang[7], Sharath Sriram[8], Kourosh Kalantar-Zadeh[9], Shuit-Tong Lee[1], Rainer Hillenbrand[4,5,10]* & Qiaoliang Bao[1,7]*

Polaritons—hybrid light–matter excitations—enable nanoscale control of light. Particularly large polariton field confinement and long lifetimes can be found in graphene and materials consisting of two-dimensional layers bound by weak van der Waals forces[1,2] (vdW materials). These polaritons can be tuned by electric fields[3,4] or by material thickness[5], leading to applications including nanolasers[6], tunable infrared and terahertz detectors[7], and molecular sensors[8]. Polaritons with anisotropic propagation along the surface of vdW materials have been predicted, caused by in-plane anisotropic structural and electronic properties[9]. In such materials, elliptic and hyperbolic in-plane polariton dispersion can be expected (for example, plasmon polaritons in black phosphorus[9]), the latter leading to an enhanced density of optical states and ray-like directional propagation along the surface. However, observation of anisotropic polariton propagation in natural materials has so far remained elusive. Here we report anisotropic polariton propagation along the surface of α-MoO_3, a natural vdW material. By infrared nano-imaging and nano-spectroscopy of semiconducting α-MoO_3 flakes and disks, we visualize and verify phonon polaritons with elliptic and hyperbolic in-plane dispersion, and with wavelengths (up to 60 times smaller than the corresponding photon wavelengths) comparable to those of graphene plasmon polaritons and boron nitride phonon polaritons[3–5]. From signal oscillations in real-space images we measure polariton amplitude lifetimes of 8 picoseconds, which is more than ten times larger than that of graphene plasmon polaritons at room temperature[10]. They are also a factor of about four larger than the best values so far reported for phonon polaritons in isotopically engineered boron nitride[11] and for graphene plasmon polaritons at low temperatures[12]. In-plane anisotropic and ultra-low-loss polaritons in vdW materials could enable directional and strong light–matter interactions, nanoscale directional energy transfer and integrated flat optics in applications ranging from biosensing to quantum nanophotonics.

Anisotropic optical materials exhibit numerous distinctive and non-intuitive optical phenomena such as negative refraction[13], hyper-lensing[14], wave-guiding[15] and enhanced quantum radiation[16], which have been demonstrated typically with artificial hyperbolic metamaterials. However, further progress is limited by optical losses and the complexity of metamaterial fabrication[17].

The recent emergence of low-loss vdW materials opens the door to achieving anisotropic optical phenomena naturally, because their layered crystal structure leads to an intrinsic and strong out-of-plane (perpendicular to the layers) optical anisotropy[5,18]. Prominent examples are hyperbolic phonon polaritons (PhPs)—infrared light coupled to lattice vibrations in layered polar materials—in hexagonal boron nitride (h-BN), which exhibit long lifetimes[11], ultra-slow propagation[19] and hyper-lensing effects[20,21]. Interestingly, when the layers of a vdW material are anisotropic (that is, when the permittivities along orthogonal in-plane directions are different), the polaritons are expected to propagate along the layers with an in-plane anisotropic dispersion[9]. When the permittivities are different but of the same sign, the polaritons possess an elliptic in-plane dispersion, in which the iso-frequency contours (slices in two-dimensional (2D) wavevector space (k_x, k_y) of constant frequency ω) describe ellipsoids. When the signs are different, the polaritons possess an in-plane hyperbolic dispersion, in which the iso-frequency contours are open hyberboloids[22]. Only recently, PhPs with in-plane hyperbolic dispersion have been demonstrated by fabricating an artificial metamaterial out of h-BN flakes[23].

Theory predicts polaritons with both in-plane anisotropies even for natural materials (without any nanostructuring) that exhibit an in-plane anisotropy of their electronic or structural properties: for example, hyperbolic plasmons—light coupled to free carriers—in black phosphorus[9] or in Weyl semimetals[24]. While being expected to provide fundamental insights into exotic material properties (for example, non-reciprocal Purcell enhancement[24]), they also bear application potential, including intrinsically non-reciprocal plasmon guiding[25], topological transitions in 2D anisotropic plasmons[22] and directional nanoscale energy collimation[26] (for use as planar and directional light emitters with on-chip integration). However, their experimental observation and verification has so far been elusive. Here we present the first (to our knowledge) images of in-plane elliptic and hyperbolic polaritons (more precisely, PhPs) that propagate with record-length lifetimes. We found them in thin slabs of α-phase molybdenum trioxide (α-MoO_3), a natural vdW polar semiconductor. Phonon polaritons have been observed[27] only recently in α-MoO_3, but their anisotropic propagation properties have not been described.

The diagrams in Fig. 1a, b show the orthorhombic crystal structure of α-MoO_3, in which layers formed by distorted MoO_6 octahedra (Fig. 1a) are weakly bound by vdW forces[28] and all three lattice constants (a, b and c) are different (Fig. 1b). Most importantly, α-MoO_3 has strong in-plane structural anisotropy, caused by the interlayer spacing of the (100) facet differing from that of the (001) facet by as much as 7.2%, which leads to the highly anisotropic response[29] (Supplementary Information). Indeed, the different directional vibrations of the α-MoO_3 crystal structure yield two infrared 'reststrahlen bands' (RBs)[30] between about 820 cm^{-1} and 1,010 cm^{-1}; in this range, the typically strong reflectivity between the transverse optical and longitudinal optical phonon frequencies (TOs and LOs, respectively) shows a large in-plane anisotropy (Supplementary Information). Thus, we can expect that in-plane anisotropic PhPs exist in this material. An optical microscopy image of the α-MoO_3 flakes and their typical Raman spectrum

[1]Institute of Functional Nano and Soft Materials (FUNSOM), Jiangsu Key Laboratory for Carbon-based Functional Materials and Devices, and Collaborative Innovation Center of Suzhou Nano Science and Technology, Soochow University, Suzhou, China. [2]Departamento de Física, Universidad de Oviedo, Oviedo, Spain. [3]Donostia International Physics Center (DIPC), Donostia-San Sebastián, Spain. [4]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain. [5]CIC nanoGUNE, Donostia-San Sebastián, Spain. [6]Department of Materials, ETH Zürich, Zürich, Switzerland. [7]Department of Materials Science and Engineering, and ARC Centre of Excellence in Future Low-Energy Electronics Technologies (FLEET), Monash University, Clayton, Victoria, Australia. [8]Functional Materials and Microsystems Research Group and MicroNano Research Facility, RMIT University, Melbourne, Australia. [9]School of Chemical Engineering, University of New South Wales (UNSW), Kensington, New South Wales, Australia. [10]CIC nanoGUNE and UPV/EHU, Donostia-San Sebastián, Spain. [11]These authors contributed equally: W. Ma, P. Alonso-González, S. Li. *e-mail: pabloalonso@uniovi.es; r.hillenbrand@nanogune.eu; qiaoliang.bao@monash.edu

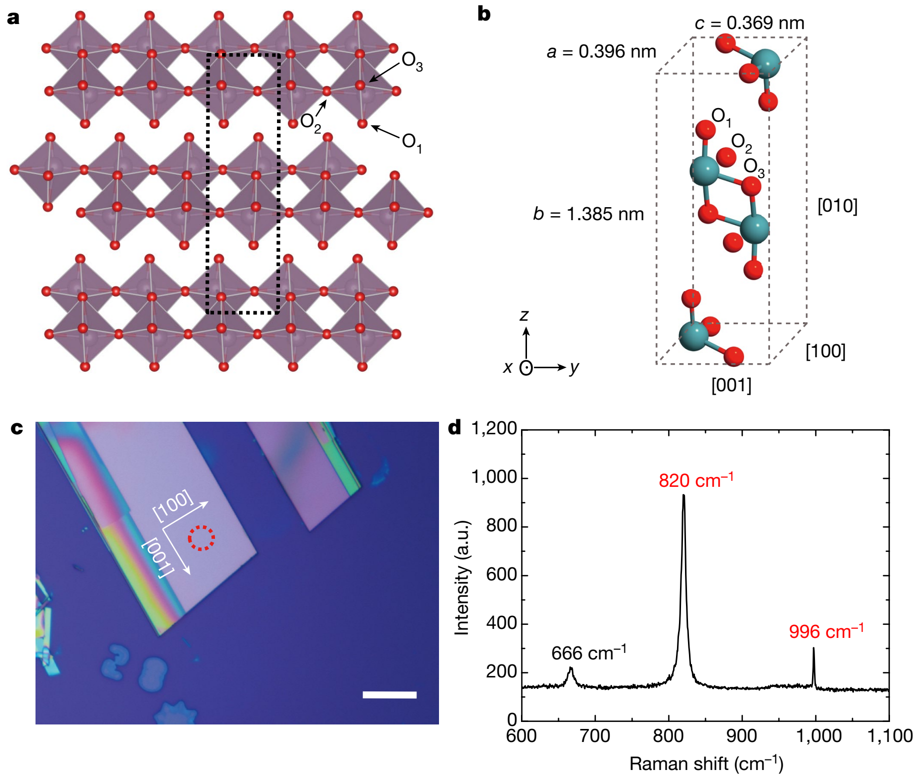


Fig. 1 | Physical properties of α-MoO_3. a, Illustration of the orthorhombic lattice structure of layered α-MoO_3 (red spheres, oxygen atoms). The orthorhombic structure is based on bilayers of distorted MoO_6 octahedra stacked along the [010] direction via vdW interactions. The three possible positions of oxygen atoms are denoted $O_{1–3}$, and the unit cell is shown dashed. **b**, Schematic of the unit cell of α-MoO_3; the lattice constants are $a = 0.396$ nm, $b = 1.385$ nm and $c = 0.369$ nm. Blue spheres, molybdenum atoms. **c**, Optical image of α-MoO_3 flakes. The α-MoO_3 crystals typically appear to be rectangular owing to the anisotropic crystal structure. Labelled arrows indicate crystal directions. Scale bar, 20 μm. **d**, Raman spectrum taken in the area marked by a red dashed circle in **c**. Red frequency labels indicate the Raman peaks associated with the lattice vibrations producing the RBs of α-MoO_3.

are shown in Fig. 1c, d, respectively. The latter shows[30] characteristic peaks at 820 cm^{-1} and 996 cm^{-1} associated with the lattice vibrations producing the RBs of α-MoO_3.

To explore the polaritonic response of α-MoO_3, we performed polariton interferometry using scattering-type scanning near-field optical microscopy (s-SNOM, Fig. 2a). A vertically oscillating metallized atomic force microscopy (AFM) tip is illuminated with p-polarized infrared light of frequency ω and field E_{inc} while scanning an α-MoO_3 flake. Acting as an infrared antenna[3–5], the tip concentrates the incident field at its very apex to give a nanoscale infrared spot for local probing of material properties and for exciting polaritons. The tip-scattered radiation is recorded simultaneously with topography, yielding nanoscale resolved near-field images (Methods). Specifically, the polaritons (described by the field E and wavelength λ) excited by the tip propagate away and are back-reflected at the flake edges, giving rise to interference fringes with a spacing $\lambda/2$.

Figure 2b shows s-SNOM near-field amplitude images of an α-MoO_3 flake with thickness $d = 250$ nm taken at $\omega = 990$ cm^{-1} and $\omega = 900$ cm^{-1}, both frequencies residing[30] inside the two RBs of α-MoO_3. For $\omega = 990$ cm^{-1} (top image) we observe bright fringes parallel to all the flake edges. They strongly resemble PhPs, similar to what has been observed in s-SNOM experiments on other polar materials[5] and recently on[27] α-MoO_3. We observe that the fringe periodicity largely depends on the propagation direction, being $\lambda_x = 950$ nm and $\lambda_y = 1{,}200$ nm for the [100] and [001] crystal directions (Supplementary Information), respectively. Apart from the deep subwavelength-scale polariton confinement $\lambda_{x,y} \ll \lambda_0 = 11.1$ μm (where λ_0 is the wavelength of the illuminating infrared light), this finding reveals a strongly anisotropic in-plane propagation (along the flake). This anisotropy becomes even more marked at $\omega = 900$ cm^{-1} (Fig. 2b lower image), where the fringes are seen only parallel to the [001] direction.

For unambiguous verification of the anisotropic polariton propagation, we recorded spectroscopic line scans[5] (Methods) along the [100] and [001] in-plane crystal directions (Fig. 2d, left and right panels, respectively). We observe two spectral bands exhibiting a series of signal maxima (fringes). The band limits (indicated by the horizontal dashed lines) correspond to the LO and TO phonon frequencies of α-MoO_3 (denoted by LO_1, LO_2, TO_1 and TO_2) and thus reveal the upper and lower RBs (U-RB and L-RB, respectively). In the U-RB we find that the fringe spacing (corresponding to the polariton wavelength) along both the [100] and [001] directions increases with increasing frequency, indicating a negative phase velocity (analogous to PhPs in the lower type-I RB of h-BN[19]). As in Fig. 2b, we observe a slightly different fringe spacing for the [100] and [001] directions, but now for all frequencies between TO_2 and LO_2. A very different behaviour is observed for the L-RB. Along the [100] direction we see fringes whose spacing decreases with increasing frequency, manifesting polaritons with positive phase velocity. More importantly, along the [001] direction we do not observe signal oscillations at a fixed frequency for the whole spectral range between TO_1 and LO_1. This finding indicates the absence of PhPs propagating in the [001] direction, supporting our assumption of a hyperbolic in-plane dispersion. The horizontal fringes observed in Fig. 2d (right panel) are caused by polaritons propagating along the [100] direction. Note that a line profile for a fixed ω corresponds to a vertical line profile (along the [001] direction, and thus parallel to the interference fringes) in the lower panel of Fig. 2b, where we can see that PhPs are launched by the left edge of the flake. Depending on ω and on the distance between the tip and the left flake edge, we thus observe either a constantly bright or dark contrast when the tip is scanned along the [001] direction, corresponding to a bright or dark horizontal fringe in the right panel of Fig. 2d.

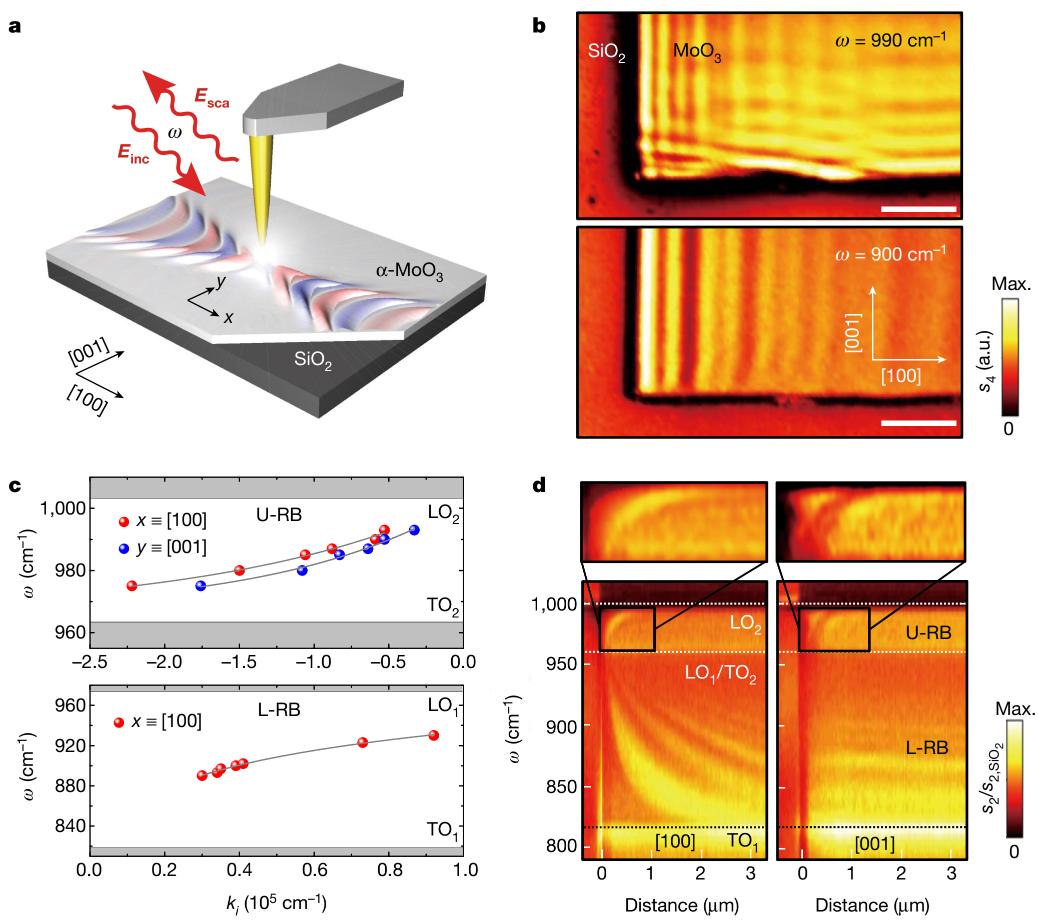


Fig. 2 | Real-space imaging and nano-spectroscopy of an α-MoO_3 flake. **a**, Schematic of the s-SNOM experimental configuration used to image an α-MoO_3 flake. A metallized AFM tip (yellow) is illuminated by p-polarized infrared light of frequency ω and electric field E_{inc}. It launches polaritons, which are back-reflected at the flake edges and subsequently scattered by the tip. The tip-scattered field E_{sca} is detected by a distant detector. **b**, Near-field amplitude images s_4 (Methods) of an α-MoO_3 flake with thickness $d = 250$ nm at illuminating frequencies $\omega = 990\ cm^{-1}$ (top panel) and $\omega = 900\ cm^{-1}$ (bottom panel). Scale bars, 2 μm. **c**, Dispersion of PhPs along the [100] and [001] directions in the U-RB (top panel) and L-RB (bottom panel, see text). Grey lines in both panels are guides to the eye. Grey shaded areas indicate the spectral regions outside the RBs. **d**, Bottom row, nano-FTIR spectral line scans along [100] and [001] (directions shown as arrows in the bottom panel of **b**), showing the near-field amplitude s_2 (normalized to the near-field amplitude on the SiO_2 substrate, s_{2,SiO_2}) as a function of distance between tip and flake edge. Dotted lines mark the approximate longitudinal and transversal phonon modes in α-MoO_3 (TO_1, 820 cm^{-1}; LO_1/TO_2, 963 cm^{-1}; LO_2, 1,003 cm^{-1}). Top row, zooms into the boxed areas of the U-RB shown in the bottom row.

For a better understanding and quantitative analysis of the anisotropic polariton propagation, we extracted the PhP dispersions, $\omega(k_i)$ ($i = x, y$), from monochromatic s-SNOM images (not shown) of the flake in Fig. 2b. The dispersions for both crystal directions and RBs are plotted in Fig. 2c. For the U-RB (upper panel), the PhP dispersions along both crystal directions are similar, although slightly separated from each other (that is, for the same frequency ω, we measured different wavevectors k_i). This result verifies that PhPs in the U-RB propagate with in-plane anisotropy. By plotting the complex-valued wavevector of the PhPs, we find that their phase velocity, $v_{p,i} = \omega/k_i$, is negative along both directions, which is indicated by negative k_i values. Furthermore, the remarkably small slopes of the dispersion curves (Supplementary Information) yield unprecedentedly small group velocities ($v_{g,i} = (\partial k_i/\partial\omega)^{-1}$) of about $0.8 \times 10^{-3}c$ (at $\omega = 985\ cm^{-1}$), which in the future could be exploited for strong light–matter interaction experiments[31]. For the L-RB (Fig. 2c lower panel), we only display the dispersion of the PhPs for the [100] direction, as no PhPs are observed in the orthogonal [001] direction. In this case, the phase velocity is positive (indicated by positive k_x values), and the group velocity is about $0.7 \times 10^{-2}c$ (at $\omega = 893\ cm^{-1}$), which is comparable to that of ultra-slow PhPs in h-BN[11].

To quantify the anisotropy of the PhPs and to measure their iso-frequency contours in wavevector space, we analyse PhP propagation along all possible directions on the flake. To that end, we fabricated disks of α-MoO_3 (Methods) and performed polariton interferometry experiments (Supplementary Information) analogous to those reported in Fig. 2. Figure 3a, b shows typical near-field amplitude images taken at frequencies respectively in the U-RB ($\omega = 983\ cm^{-1}$) and in the L-RB ($\omega = 893\ cm^{-1}$) of α-MoO_3. In the U-RB, the interference pattern shows an elliptical shape with the largest PhP wavelength along the [001] surface direction, which continuously reduces to its smallest value along the orthogonal [100] surface direction. More strikingly, in the L-RB the interference pattern manifests as an almond shape, in which the PhPs have their largest wavelength along the [100] direction and which continuously reduces to zero until no discernible polariton propagation occurs along the orthogonal [001] direction. By Fourier transform of Fig. 3a, b, we obtain the iso-frequency contours directly. We find an ellipsoid in the U-RB (Fig. 3c) and a hyperbola in the L-RB (Fig. 3d), revealing that the PhPs exhibit elliptic and hyperbolic dispersions, respectively. Note that Fig. 3c shows two ellipses instead of one, differing by a factor of 2 in their semi-axes. We attribute this observation to the presence of both tip- and edge-launched PhPs[32,33] in Fig. 3a (Supplementary Information). On the other hand, the hyperbola in Fig. 3d opens along the [001] direction, which indicates that PhPs along this crystal direction are forbidden, thus explaining the observations in Figs. 2, 3b.

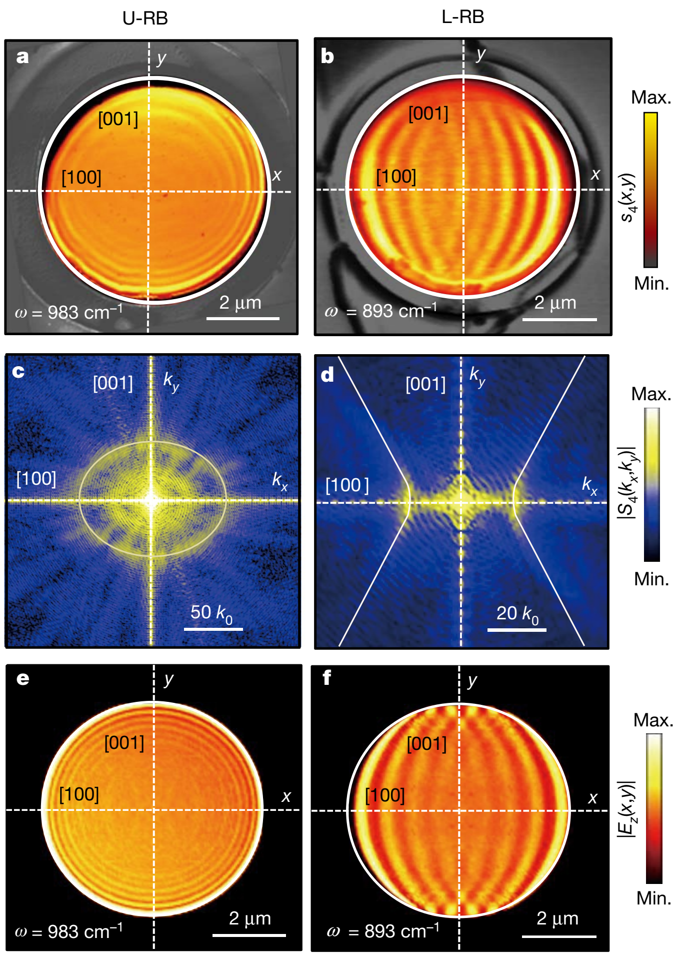


Fig. 3 | In-plane elliptical and hyperbolic PhPs in an α-MoO_3 disk. **a**, **b**, Near-field amplitude images s_4 of an α-MoO_3 disk (colour key at right) with $d = 144$ nm. The imaging frequencies are $\omega = 983$ cm^{-1} (U-RB; **a**), and 893 cm^{-1} (L-RB; **b**). Dashed white lines indicate the [100] and [001] surface directions. Scale bars, 2 µm. **c**, **d**, Absolute value of the Fourier transform $|S_4(k_x,k_y)|$ of the near-field images in **a** and **b**, respectively (colour key at right), revealing the iso-frequency contours for each RB. Solid lines show the iso-frequency contours of the PhPs obtained by fitting equation (1) for each case (note that they correspond to $2k$). Scale bars are $50k_0$ and $20k_0$, for the U-RB and L-RB respectively, with k_0 being the momentum of light in free space. **e**, **f**, Calculated near-field amplitude images $|E_z(x,y)|$ (Supplementary Information) for an α-MoO_3 disk (colour key at right) at $\omega = 983$ cm^{-1} (U-RB; **e**) and 893 cm^{-1} (L-RB; **f**). Scale bars, 2 µm.

To corroborate our experimental results theoretically, and to extract the as-yet unknown anisotropic permittivity of α-MoO_3, we model the α-MoO_3 flake as a 2D conductivity layer of zero thickness (Methods). We find the following dispersion relation for polaritons in a thin in-plane anisotropic slab surrounded by two dielectric half-spaces with isotropic permittivities ε_1 and ε_2 (Supplementary Information):

$$\left[k_x^2\alpha_{xx}+k_y^2\alpha_{yy}+\frac{k_0k_t^2}{2}\left(\frac{\varepsilon_1}{k_{z1}}+\frac{\varepsilon_2}{k_{z2}}\right)\right]\left[k_y^2\alpha_{xx}+k_x^2\alpha_{yy}+\frac{k_t^2}{2k_0}(k_{z1}+k_{z2})\right]-k_x^2k_y^2(\alpha_{xx}-\alpha_{yy})^2=0 \quad (1)$$

where $k_{x,y}$ and $k_{z1,2}=\sqrt{\varepsilon_{1,2}k_0^2-k_x^2-k_y^2}$ are the in- and out-of-plane wavevectors, respectively, $k_0 = 2\pi/\lambda_0$ is the wavevector in free space, and $\hat{\alpha} = 2\pi\hat{\sigma}_{\rm eff}/c$ is the normalized conductivity, introduced for convenience. Using equation (1) with α_{xx} and α_{yy} as fitting parameters, we obtain excellent agreement with the elliptical and hyperbolic features in Fig. 3c, d, where our fits are shown as white solid lines. Neglecting absorption, we find $\alpha_{xx} = -0.12i$ ($\varepsilon_{xx} = 2.6$) and $\alpha_{yy} = -0.16i$ ($\varepsilon_{yy} = 3.7$) for $\omega = 983$ cm^{-1} (U-RB), and $\alpha_{xx} = 0.26i$ ($\varepsilon_{xx} = -6.4$) and $\alpha_{yy} = -0.07i$ ($\varepsilon_{yy} = 1.7$) for $\omega = 893$ cm^{-1} (L-RB).

We corroborate the model and permittivity values by numerical simulation of near-field images of an α-MoO_3 disk on SiO_2 (Supplementary Information). We used the nominal experimental values of 144 nm and 6 µm for the disk thickness and diameter, respectively, and the anisotropic real-valued permittivities obtained from the fit described above. The imaginary parts of the permittivities and the value of ε_{zz} (not obtained from the fit) were adjusted to obtain the best matching of the experimental images and of the sign of the phase velocities in each RB. As a result of our analysis, we find $\varepsilon_{xx} > 0$, $\varepsilon_{yy} > 0$ and $\varepsilon_{zz} < 0$ for the elliptic U-RB, and $\varepsilon_{xx} < 0$, $\varepsilon_{yy} > 0$ and $\varepsilon_{zz} > 0$ for the hyperbolic L-RB (Supplementary Information). The simulated polariton interferometry amplitude images are shown in Fig. 3e, f. Their excellent agreement with the experiments (Fig. 3a, b) validates both the model and the permittivity values. The results demonstrate that experimental PhP interferometry of α-MoO_3 disks and fitting of the results with our simple theoretical model allow the highly anisotropic local permittivities of α-MoO_3 to be extracted.

The conductivity tensor $\hat{\sigma}_{\rm eff}$—and thus the wavevectors of the PhPs—depends on the slab thickness d. According to the relation between $\hat{\sigma}_{\rm eff}$ and $\hat{\varepsilon}$ (see Methods), we obtain from equation (1) the thickness-dependent anisotropic in-plane polariton wavevectors (Supplementary Information):

$$k_i \approx -\frac{\varepsilon_1+\varepsilon_2}{d\varepsilon_{ii}}, i = x, y \quad (2)$$

In Fig. 4a, we demonstrate the thickness tunability of in-plane hyperbolic polaritons: we plot the PhP dispersions obtained by s-SNOM nano-imaging along the [100] crystal direction of α-MoO_3 flakes with different thickness d. We clearly observe that the wavevector k_x and thus the polariton confinement increase with decreasing thickness. For $d = 55$ nm, we find k_x values of about 3.5×10^5 cm^{-1}, corresponding to a PhP wavelength of 180 nm. This value is 60 times smaller than $\lambda_0 = 10.8$ µm, suggesting that in-plane anisotropic propagation could be well paired with deep subwavelength-scale field confinement for the development of ultra-compact devices. The inverse dependence of k_x on d is better observed in Fig. 4b, where we plot the experimental k_x (red dots) obtained at $\omega = 902$ cm^{-1} for four flakes of different thickness. These experimental values are well matched by our equation (2) (grey curve), where we used $\varepsilon_{xx} = -5.1$ as extracted for the flake with $d = 144$ nm in Fig. 3, thus strongly supporting the validity of our approximation.

A key property of polaritons for future applications is their lifetime[10,11]. To measure it, we fitted s-SNOM amplitude line profiles along the [100] direction (blue and red crosses in Fig. 4c) with an exponentially decaying sine-wave function corrected by the geometrical spreading factor $\sqrt{x}$ (Supplementary Information)[10]. From the amplitude decay length L_x (one of the fitting parameters) we obtain the lifetime according to $\tau_x = L_x/v_g$, where the group velocity v_g is taken from Fig. 2c. For the in-plane hyperbolic PhPs we obtain $\tau_x = 1.9 \pm 0.3$ ps, which reveals the ultra-low-loss character of these polaritons. Surprisingly, for the in-plane elliptic PhPs we obtain $\tau_x = 8 \pm 1$ ps (four times higher than that of PhPs in isotropically enriched h-BN[11]). On some flakes we find lifetimes up to 22 ps (Supplementary Information). We note that in contrast to low-loss h-BN PhPs[11] and graphene plasmons[12], a rather small number of fringes were observed on α-MoO_3 flakes. This can be explained by the small group velocities of the MoO_3 PhPs, which yield relatively short propagation lengths. The ultra-long PhP lifetimes are corroborated by the ultra-narrow linewidths of the α-MoO_3 Raman peaks (Supplementary Information) at 996 cm^{-1} and 820 cm^{-1} (corresponding to anisotropic bond stretching modes[30] that originate in the U-RB and L-RB, respectively), revealing very high crystal quality. A similar relation has been recently reported to explain the large lifetimes observed in isotopically enriched h-BN[11].

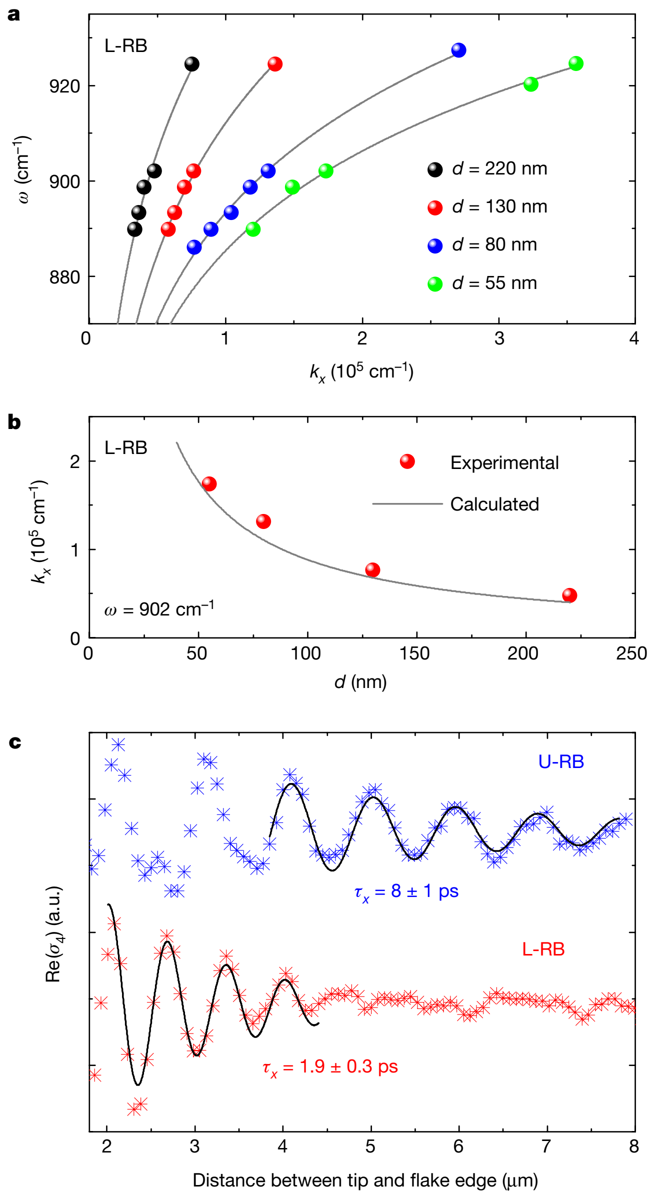


Fig. 4 | Thickness tunability and lifetime of in-plane hyperbolic and elliptic PhPs in α-MoO_3. a, Experimental (dots) PhP dispersions along the [100] direction in α-MoO_3 for a varying flake thickness d (lines are guides to the eyes). **b**, Experimental (dots) and calculated (line) dependence of k_x upon d. **c**, s-SNOM line traces (showing the real part of the complex-valued s-SNOM signal σ_4; Methods) along the [100] direction of the flake shown in Fig. 2b with $d = 250$ nm in the elliptic (blue crosses, $\omega = 990$ cm^{-1}) and hyperbolic (red crosses, $\omega = 930$ cm^{-1}) regimes. Damped sine-wave functions (black solid lines) were fitted to the data that correspond to edge-launched PhPs (Supplementary Information). Amplitude lifetimes $\tau_x = 8 \pm 1$ ps and $\tau_x = 1.9 \pm 0.3$ ps are obtained for the U-RB and L-RB, respectively.

In-plane anisotropic α-MoO_3 PhPs add a new member to the growing list of polaritons in vdW materials. In combination with external stimuli, such as strain, electric gating or photo-injection of carriers, we envisage active tuning of the anisotropic PhP properties. Our findings may thus establish a route to directional control of light and light–matter interactions at the nanoscale.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0618-9.

Received: 8 April 2018; Accepted: 17 August 2018;
Published online 24 October 2018.

1. Basov, D., Fogler, M. & de Abajo, F. G. Polaritons in van der Waals materials. *Science* **354**, aag1992 (2016).
2. Low, T. et al. Polaritons in layered two-dimensional materials. *Nat. Mater.* **16**, 182–194 (2017).
3. Fei, Z. et al. Gate-tuning of graphene plasmons revealed by infrared nano-imaging. *Nature* **487**, 82–85 (2012).
4. Chen, J. et al. Optical nano-imaging of gate-tunable graphene plasmons. *Nature* **487**, 77–81 (2012).
5. Dai, S. et al. Tunable phonon polaritons in atomically thin van der Waals crystals of boron nitride. *Science* **343**, 1125–1129 (2014).
6. Chakraborty, S. et al. Gain modulation by graphene plasmons in aperiodic lattice lasers. *Science* **351**, 246 (2016).
7. Cai, X. et al. Plasmon-enhanced terahertz photodetection in graphene. *Nano Lett.* **15**, 4295–4302 (2015).
8. Rodrigo, D. et al. Mid-infrared plasmonic biosensing with graphene. *Science* **349**, 165–168 (2015).
9. Low, T. et al. Plasmons and screening in monolayer and multilayer black phosphorus. *Phys. Rev. Lett.* **113**, 106802 (2014).
10. Woessner, A. et al. Highly confined low-loss plasmons in graphene–boron nitride heterostructures. *Nat. Mater.* **14**, 421–425 (2015).
11. Giles, A. J. et al. Ultralow-loss polaritons in isotopically pure boron nitride. *Nat. Mater.* **17**, 134–139 (2018).
12. Ni, G. X. et al. Fundamental limits to graphene plasmonics. *Nature* **557**, 530–533 (2018).
13. Hoffman, A. J. et al. Negative refraction in semiconductor metamaterials. *Nat. Mater.* **6**, 946–950 (2007).
14. Liu, Z., Lee, H., Xiong, Y., Sun, C. & Zhang, X. Far-field optical hyperlens magnifying sub-diffraction-limited objects. *Science* **315**, 1686 (2007).
15. Podolskiy, V. A. & Narimanov, E. E. Strongly anisotropic waveguide as a nonmagnetic left-handed system. *Phys. Rev. B* **71**, 201101 (2005).
16. Cortes, C. L., Newman, W., Molesky, S. & Jacob, Z. Quantum nanophotonics using hyperbolic metamaterials. *J. Opt.* **14**, 063001 (2012).
17. Takayama, O., Bogdanov, A. A. & Lavrinenko, A. V. Photonic surface waves on metamaterial interfaces. *J. Phys. Condens. Matter* **29**, 463001 (2017).
18. Caldwell, J. D. et al. Sub-diffractional volume-confined polaritons in the natural hyperbolic material hexagonal boron nitride. *Nat. Commun.* **5**, 5221 (2014).
19. Yoxall, E. et al. Direct observation of ultraslow hyperbolic polariton propagation with negative phase velocity. *Nat. Photon.* **9**, 674–678 (2015).
20. Li, P. et al. Hyperbolic phonon-polaritons in boron nitride for near-field optical imaging and focusing. *Nat. Commun.* **6**, 7507 (2015).
21. Dai, S. et al. Subdiffractional focusing and guiding of polaritonic rays in a natural hyperbolic material. *Nat. Commun.* **6**, 6963 (2015).
22. Gomez-Diaz, J. S., Tymchenko, M. & Alù, A. Hyperbolic plasmons and topological transitions over uniaxial metasurfaces. *Phys. Rev. Lett.* **114**, 233901 (2015).
23. Li, P. et al. Infrared hyperbolic metasurface based on nanostructured van der Waals materials. *Science* **359**, 892–896 (2018).
24. Song, J. C. W. & Rudner, M. S. Fermi arc plasmons in Weyl semimetals. *Phys. Rev. B* **96**, 205443 (2017).
25. Mazor, Y. & Steinberg, B. Z. Longitudinal chirality, enhanced nonreciprocity, and nanoscale planar one-way plasmonic guiding. *Phys. Rev. B* **86**, 045120 (2012).
26. Kildishev, A. V., Boltasseva, A. & Shalaev, V. M. Planar photonics with metasurfaces. *Science* **339**, 1232009 (2013).
27. Zheng, Z. et al. Highly confined and tunable hyperbolic phonon polaritons in van der Waals semiconducting transition metal oxides. *Adv. Mater.* **30**, 1705318 (2018).
28. de Castro, I. A. et al. Molybdenum oxides — from fundamentals to functionality. *Adv. Mater.* **29**, 1701619 (2017).
29. Lajaunie, L., Boucher, F., Dessapt, R. & Moreau, P. Strong anisotropic influence of local-field effects on the dielectric response of α-MoO_3. *Phys. Rev. B* **88**, 115141 (2013).
30. Py, M. A., Schmid, P. E. & Vallin, J. T. Raman scattering and structural properties of MoO_3. *Nuovo Cimento B* **38**, 271–279 (1977).
31. Caldwell, J. D. et al. Low-loss, infrared and terahertz nanophotonics using surface phonon polaritons. *Nanophotonics* **4**, 44–68 (2015).
32. Dai, S. et al. Efficiency of launching highly confined polaritons by infrared light incident on a hyperbolic material. *Nano Lett.* **17**, 5285–5290 (2017).
33. Hu, F. et al. Imaging the localized plasmon resonance modes in graphene nanoribbons. *Nano Lett.* **17**, 5423–5428 (2017).

Acknowledgements We thank S. C. Dhanabalan and J. S. Ponraj for their efforts in the early stages of this project. We thank M. H. Lu, L. Liu, C. W. Qiu and L. Wang for suggestions, and H. Yan and Q. Xing for their assistance with micro-FTIR measurements. We thank Quantum Design China (Beijing laboratory) for technical support of some s-SNOM measurements. This work was performed in part at the Melbourne Centre for Nanofabrication (MCN) in the Victorian Node of the Australian National Fabrication Facility (ANFF). We acknowledge support from the National Natural Science Foundation of China (grant numbers 51222208, 51290273, 51601131, 61604102, 51702219 and 91433107), the Youth 973 programme (2015CB932700), the National Key Research and Development Program (2016YFA0201900), ARC (DP140101501, IH150100006, FT150100450 and CE170100039),

the Natural Science Foundation of Jiangsu Province (BK20150053), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), the Collaborative Innovation Center of Suzhou Nano Science and Technology, and the Spanish Ministry of Economy, Industry and Competitiveness (national projects MAT2015-65525-R, FIS2014-60195-JIN, MAT2017-88358-C3-3-R, MAT2014-53432-C5-4-R, and the project MDM-2016-0618 of the Maria de Maeztu Units of Excellence Programme). Q.B. acknowledges support from the Australian Research Council (ARC) Centre of Excellence in Future Low-Energy Electronics Technologies (FLEET). P.A.-G. acknowledges support from the European Research Council under Starting Grant 715496, 2DNANOPTICA. J.M.-S. acknowledges support through the Clarín Programme from the Government of the Principality of Asturias and a Marie Curie-COFUND grant (PA-18-ACB17-29). P.L. acknowledges support from a Marie Sklodowska-Cuire individual fellowship (SGPCM-705960).

Reviewer information *Nature* thanks A. Chaves and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions W.M., P.A.-G. and S.L. contributed equally to this work. Q.B. conceived the initial measurements on α-MoO_3. R.H., P.A.-G. and Q.B. supervised the project. W.M. and P.A.-G. carried out the near-field imaging experiments with the help of I.A., J.M.-S., J.T.-G., Z.D. and P.L. J.Y. carried out the far-field experiments. W.M., P.A.-G., A.Y.N., S.L., R.H. and Q.B. participated in data analysis and co-wrote the manuscript. A.Y.N. suggested the model and supervised the development of the theory. J.M.-S., J.T.-G. and P.A.-G. carried out the simulations. Y.J., S.S., Y.Z. and K.K.-Z. contributed to the material synthesis. S.V., C.T., Z.D. and Y.Z. contributed to sample fabrication.

Competing interests R.H. is cofounder of Neaspec GmbH, a company producing scattering-type near-field scanning optical microscope systems, such as the one used in this study. The remaining authors declare no competing financial interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0618-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to P.A. or R.H. or Q.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

s-SNOM and nano-FTIR set-up. For infrared nano-imaging we used a scattering-type scanning near-field optical microscope (s-SNOM[3–5], from Neaspec). Metallized, cantilevered atomic force microscope (AFM) tips are used as scattering near-field probes. The tip is oscillating vertically at the mechanical resonant frequency (around 270 kHz) of the cantilever, with an amplitude of about 50 nm. The tip is illuminated with p-polarized infrared light of frequency ω (from tunable CO_2 and quantum cascade lasers) and electric E_{inc}, while the α-MoO_3 flake is raster-scanned below the oscillating tip. Acting as an infrared antenna, the Pt-coated tip concentrates the incident field into a nanoscale spot at the apex, which interacts with the sample surface and thus modifies the tip-scattered field E_{sca}. E_{sca} is recorded with a pseudo-heterodyne Michelson interferometer[34]. Demodulation of the interferometric detector signal at the nth harmonics of the tip oscillation frequency yields the complex-valued near-field signals $\sigma_n = s_n e^{i\varphi_n}$, with s_n being the near-field amplitude and φ_n being the near-field phase. By recording the near-field signals as a function of the lateral tip position, we obtain near-field images or line trace. In the particular case of probing a material supporting polaritons, the nanoscale 'hotspot' at the tip apex acts as a local source of polaritons[3–5]. The tip-launched polaritons reflect at the flake edges and produce polariton interference, yielding fringes in the near-field images (Figs. 2b, 3a, b). The distance between the interference fringes corresponds to half the polariton wavelength, $\lambda/2$.

For nano-FTIR spectroscopy[35], the tip was illuminated by a broadband super-continuum laser, and the tip-scattered light was recorded with an asymmetric Fourier transform spectrometer. By recording point spectra as a function of the tip position, we obtained high-resolution spectral line scans[5].

Disk fabrication. Bulk MoO_3 crystals were grown via chemical vapour deposition. Commercial MoO_3 powder (Sigma-Aldrich) was evaporated in a horizontal tube furnace at 785 °C and was re-deposited as α-MoO_3 crystals at 560 °C. The deposition process was carried out in an inert environment (Ar flow of 200 sccm) at 1 torr[28]. The as-grown bulk crystals were then mechanically exfoliated and transferred onto a Si/SiO_2 (thickness 300 nm) substrate. The transferred flakes were inspected with an optical microscope and characterized via AFM, allowing the selection of large and homogeneous pieces with the desired thickness. The selected flake was then shaped into a disk by using focused Ga-ion beam milling in a FEI Helios 600 Nanolab dual beam system. In order to protect the surface of the disk from the implantation of Ga ions, the flake was first covered by placing a thin diamond shield (approximate dimensions $100\,\mu m \times 80\,\mu m \times 0.5\,\mu m$) on top of the flake using the tip of an Omniprobe micromanipulator. Using the ion beam, we then milled through both the diamond shield and the flake using a ring-shaped milling pattern, until we reached the substrate. The remaining parts of the diamond shield were then lifted off the surface using the Omniprobe micromanipulator to give only the disk-shaped flake separated from the bulk flake by a ring-shaped channel.

Conductivity model for MoO_3 layers. Modelling the α-MoO_3 flake as a 2D conductivity layer of zero thickness avoids the calculation of the fields inside the slab[22], and has been proven valid for in-plane isotropic 2D materials (for example, graphene[36] and transition layer polaritons[37]) with a layer thickness that is much smaller than the polariton wavelength ($d \ll \lambda$). In the model, the effective conductivity for the isotropic layer is given by $\sigma_{eff} = [cd/(2i\lambda_0)]\varepsilon$, where ε is the in-plane isotropic permittivity (both ε and σ_{eff} are scalars). Note that σ_{eff} scales linearly with d, thus taking into account the effect of the small slab thickness. Analogously, we model the α-MoO_3 layer by an anisotropic in-plane conducting layer with zero thickness and an effective 2D conductivity tensor, $\hat{\sigma}_{eff}$. The generalized relation between the tensor $\hat{\sigma}_{eff}$ and the (2×2) permittivity tensor $\hat{\varepsilon} = \mathrm{diag}(\varepsilon_{xx}, \varepsilon_{yy})$ is then given by $\hat{\sigma}_{eff} = (cd/2i\lambda_0)\hat{\varepsilon}$. Note that the model is independent of the out-of-plane permittivity component ε_{zz}, which subsequently does not enter into equation (1).

Data availability

All the data are available in the online version of the paper. The data that support the findings of this study are available from the corresponding authors on reasonable request.

34. Ocelic, N., Huber, A. & Hillenbrand, R. Pseudoheterodyne detection for background-free near-field spectroscopy. *Appl. Phys. Lett.* **89**, 101124 (2006).
35. Huth, F., Schnell, M., Wittborn, J., Ocelic, N. & Hillenbrand, R. Infrared-spectroscopic nanoimaging with a thermal source. *Nat. Mater.* **10**, 352–356 (2011).
36. Nikitin, A. Y. in *World Scientific Handbook of Metamaterials and Plasmonics* 307–338 (World Scientific Series in Nanoscience and Nanotechnology, World Scientific, Singapore, 2017).
37. Tilley, D. R. *Surface Polaritons: Electromagnetic Waves at Surfaces and Interfaces* (North-Holland Publishing Co., Amsterdam, 1982).

https://doi.org/10.1038/s41586-018-0608-y

A protein functionalization platform based on selective reactions at methionine residues

Michael T. Taylor[1], Jennifer E. Nelson[1], Marcos G. Suero[1] & Matthew J. Gaunt[1]*

Nature has a remarkable ability to carry out site-selective post-translational modification of proteins, therefore enabling a marked increase in their functional diversity[1]. Inspired by this, chemical tools have been developed for the synthetic manipulation of protein structure and function, and have become essential to the continued advancement of chemical biology, molecular biology and medicine. However, the number of chemical transformations that are suitable for effective protein functionalization is limited, because the stringent demands inherent to biological systems preclude the applicability of many potential processes[2]. These chemical transformations often need to be selective at a single site on a protein, proceed with very fast reaction rates, operate under biologically ambient conditions and should provide homogeneous products with near-perfect conversion[2–7]. Although many bioconjugation methods exist at cysteine, lysine and tyrosine, a method targeting a less-explored amino acid would considerably expand the protein functionalization toolbox. Here we report the development of a multifaceted approach to protein functionalization based on chemoselective labelling at methionine residues. By exploiting the electrophilic reactivity of a bespoke hypervalent iodine reagent, the *S*-Me group in the side chain of methionine can be targeted. The bioconjugation reaction is fast, selective, operates at low-micromolar concentrations and is complementary to existing bioconjugation strategies. Moreover, it produces a protein conjugate that is itself a high-energy intermediate with reactive properties and can serve as a platform for the development of secondary, visible-light-mediated bioorthogonal protein functionalization processes. The merger of these approaches provides a versatile platform for the development of distinct transformations that deliver information-rich protein conjugates directly from the native biomacromolecules.

The sheer structural diversity of the proteome in any single organism means that no one protein functionalization method is likely to provide a universal solution for the preparation of protein constructs[8–11] (Fig. 1a). Although encoded by the AUG start codon at the beginning of protein synthesis, methionine is often post-translationally excised and thus has a low abundance in proteins (around 2%). It is also frequently used as a replacement for hydrocarbon-containing residues[12]. The limited function of methionine, being mainly responsible for protection against oxidative stress, compared to other residues means that its functionalization is less likely to impair protein function[13]. Targeting methionine would not only provide a distinct bioconjugation approach but also, using our strategy, the resulting methionine conjugates would provide exploitable intrinsic reactivity such that they could lead to the rapid synthesis of diverse, functional protein constructs from native proteins.

So far, there has been only one effective method reported for bioconjugation at methionine, in which the oxidation of thioethers with oxaziridine reagents provided the basis for an elegant bio-inspired strategy to form stable protein-bound sulfoximines[14]. We aimed to target the polarizable thioether on the methionine side chain with a suitable electrophile to form a cationic sulfonium species, selectively installing a versatile payload and distinct functionality at a methionine residue, thereby providing a fundamentally different bioconjugation approach[8–11] (Fig. 1a). Methionine residues react with cyanogen bromide (CNBr); however, such a process cannot function as a bioconjugation method because the instability of the resulting cyano-substituted sulfonium cation triggers cleavage of the protein backbone[15]. Conversely, methionine residues undergo slow *S*-alkylation reactions with iodoacetamide or with other benzyl-derived electrophiles to form relatively stable trialkyl sulfonium cations[16–19]. The need to strike a balance between the stability of the protein-bound sulfonium cations, the compatibility of the reaction conditions, and the reaction rate of the thioether with suitable electrophiles has, so far, precluded the development of an effective alkylation-based method for bioconjugation at methionine. Guided by these limitations, we posited that a distinct class of electrophile, based on the hypervalent iodine scaffold of λ^3-iodanes[20], could make for a functioning bioconjugation process at methionine. Tailoring the substituents and counteranion on the I(III) atom should enable us to tune both the reactivity of the polarizable I(III) nodal centre, to dovetail with the electron lone pair of the thioether, and also the stability of the resulting sulfonium conjugate, through modulation of the electronic features of the groups directly attached to the cationic sulfur motif. We noted that a structurally remarkable iodonium salt (**1** in Fig. 1, R = Et and X = OTf (trifluoromethanesulfonate, also known as triflate)) reacts rapidly with dimethylsulfide (the simplest possible mimic of the thioether motif in methionine) to form a sulfonium adduct[21,22]. Successful reaction of this iodonium salt with methionine would not only represent a distinct method for bioconjugation, but also deliver a high-energy conjugate equipped with reactive 'on-protein' groups that could serve as a basis for designing new transformations towards protein constructs with diverse functionality (Fig. 1b).

We first examined the reaction between dipeptide **2a** and iodonium triflate **1a** (Fig. 2a). We observed the formation of the desired sulfonium conjugate **3a**, although it was clear from the low yield (27%) that the iodonium salt was poorly stable in aqueous solution. By tailoring the aryl group of the iodonium salt (to the electron-deficient 2,4-difluorobenzene) and replacing the triflate counteranion with tetrafluoroborate, we found that a readily prepared reagent **1b** displayed superior physical properties (half-life in water is >50 h). Treatment of reagent **1b** with dipeptide **2a** gave 72% of the desired product **3b** accompanied by the corresponding sulfoxide (not shown) after reaction for 30 min.

Moving to a more complex substrate, the GLP-1 receptor agonist exenatide (Byetta, **2b**, a 39-residue helical polypeptide containing a single mid-chain methionine, Fig. 2b), we found that treatment of a 100 μM aqueous solution with **1b** led to decomposition of the polypeptide. A key breakthrough revealed that the addition of a low concentration of thiourea (20 mM) resulted in a substantial improvement of the reaction, such that sulfonium conjugate **4a** was formed with 68% conversion in less than 2 min, accompanied by non-specific oxidation and labelling. Further improvements could be made by performing the reaction in the presence of TEMPO ((2,2,6,6-tetramethylpiperidin-1-yl)oxyl, 10 mM), which minimized

[1]Department of Chemistry, University of Cambridge, Cambridge, UK. *e-mail: mjg32@cam.ac.uk

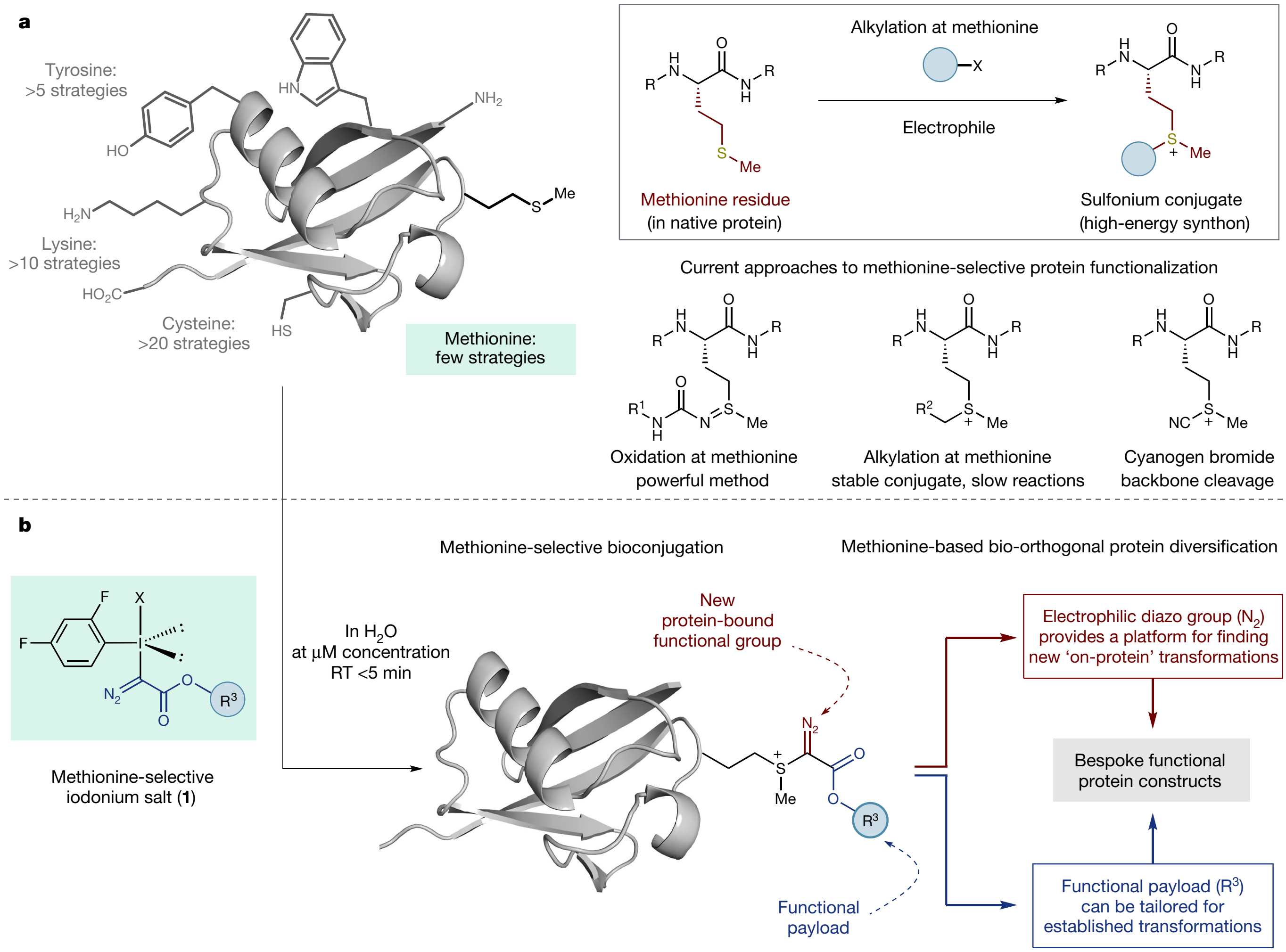


Fig. 1 | The development of a methionine-selective protein functionalization strategy. a, Existing protein functionalization strategy, and the potential for methionine-selective bioconjugation. R = peptide or protein; R^1 = various organic groups; R^2 = aryl or ester group. **b**, Functionalized hypervalent iodine reagents enable methionine-selective protein modification, leading to methionine-based bioorthogonal protein diversification. X = leaving group; R^3 = functional payload.

the formation of oxidative by-products, and adding aqueous formic acid solution (5 mM, approximately pH 3), which reduced the formation of non-specifically labelled by-products to trace levels. Finally, we found that the labelling process proceeded effectively when conducted in distilled water. Routine analysis of electrospray ionization mass spectrometry (ESI-MS) and tandem mass spectrometry (MS/MS) fragmentation data confirmed selective reaction at methionine and, although the concentration of thiourea is well below the levels needed for protein denaturation, it was confirmed by circular dichroism spectra of exenatide conjugate **4a** that the characteristic helical structure was retained (see Supplementary Fig. 10). Although we are not yet certain of the role of thiourea, it is important to note that its presence appears to be fundamental in providing a bioconjugation process at reaction rates needed for transformation on complex biomolecules. With these refined conditions, the thiourea-accelerated bioconjugation was almost instantaneous at exenatide concentrations ranging from

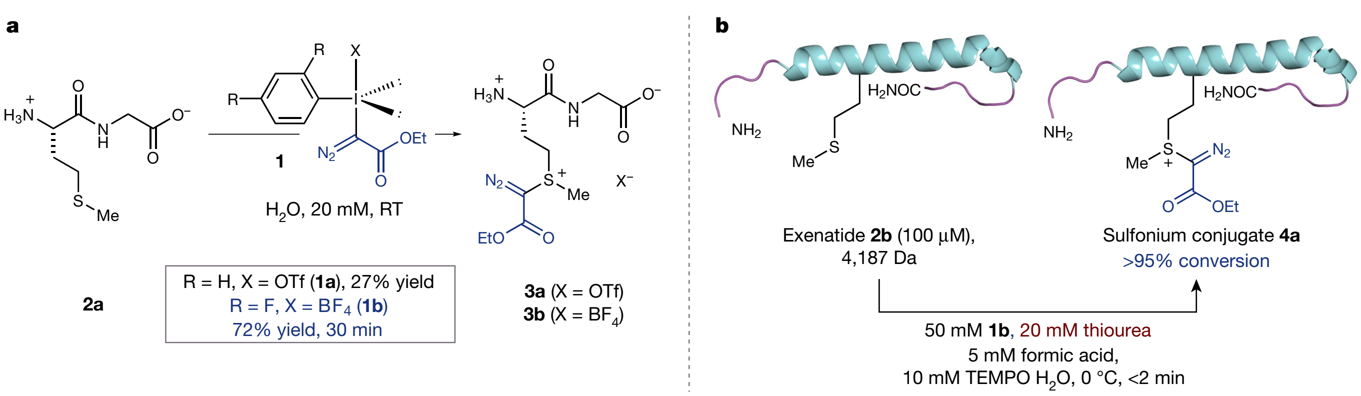


Fig. 2 | Evolution of a methionine-selective bioconjugation strategy. a, Initial results for functionalization at methionine with hypervalent iodine reagents. **b**, Optimal process for the thiourea-accelerated methionine-selective bioconjugation of exenatide.

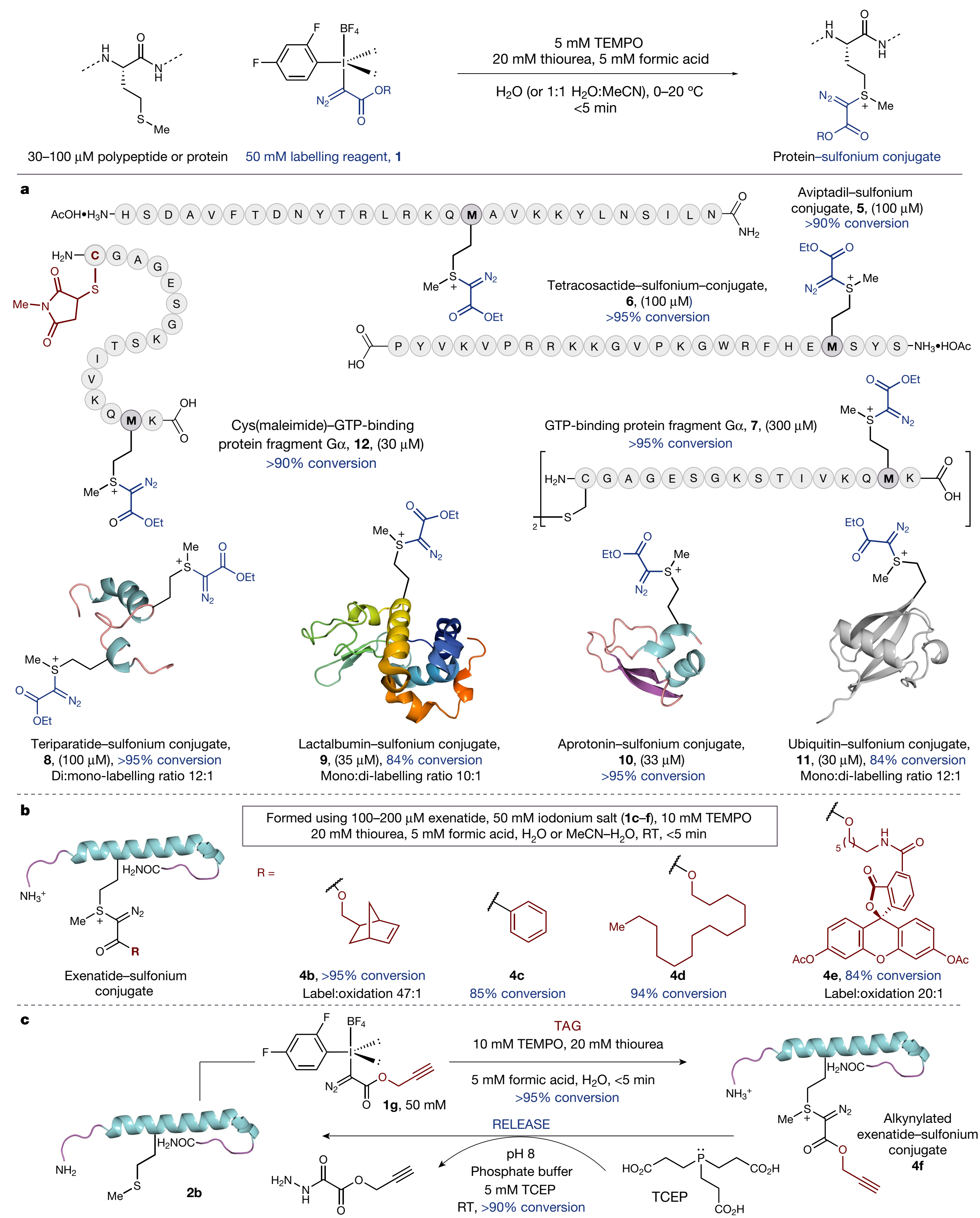


Fig. 3 | Scope of the methionine-selective bioconjugation strategy. **a**, Range of polypeptide or protein substrates. **b**, Functionalized iodonium reagents compatible with the bioconjugation. **c**, A stimuli-responsive strategy for reversing methionine bioconjugation.

5–500 μM without compromising the conversion, and a larger, milligram-scale reaction enabled the purification of conjugate **4a** by semi-preparative high-performance liquid chromatography to give a 79% yield of isolated product. Given the similarity of the exenatide conjugate to the intermediate that would arise from reaction of the polypeptide with cyanogen bromide (leading to peptide cleavage),

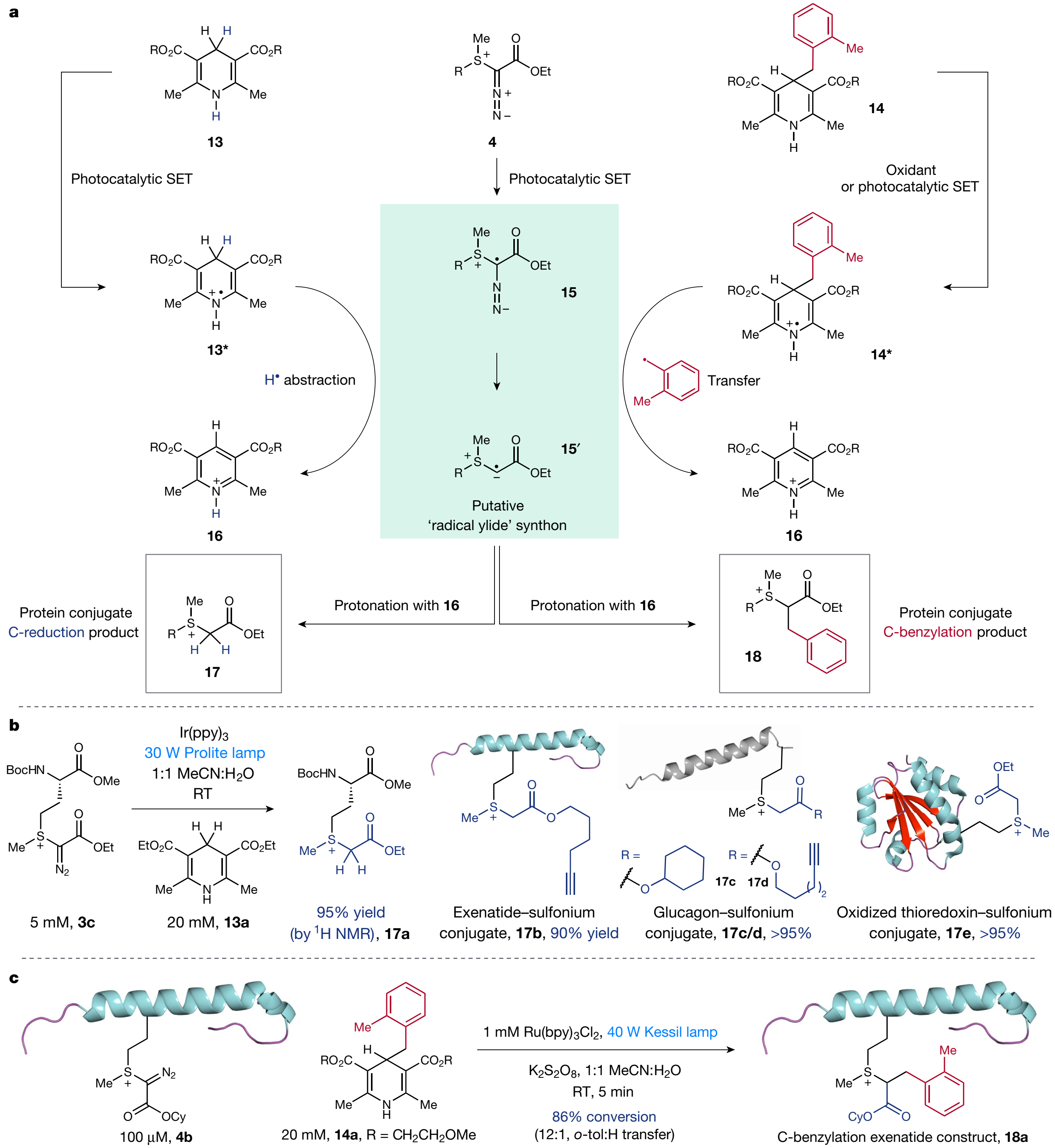


Fig. 4 | Exploiting the multi-faceted reactivity of the protein–sulfonium conjugate. a, A photocatalytic design plan for secondary protein diversification. SET, single-electron transfer. **b**, A system for photoredox-mediated reduction of the sulfonium conjugate and examples of the substrate scope. Notably, the reduction of **3c** proceeds in light, without the iridium photocatalyst, but the yield is greatly reduced. **c**, Secondary protein functionalization via photoredox radical cross-coupling between the diazo sulfonium–protein conjugate and the C-4 substituted Hantzsch ester.

it is remarkable that **4a** has a half-life in water of over 100 h. We believe that the observed stability of **4a** is a result of the ethyl diazoacetate motif imparting a lesser electron-withdrawing effect (than the cyano group) on the sulfonium salt, which in turn leads to a species that is less reactive to attack by the proximal carbonyl group and is, therefore, a more stable conjugate.

The optimized conditions, using **1b**, were used to evaluate the substrate scope (Fig. 3a). Random coil polypeptides that contain methionine, such as aviptadil and tetracosactide, are efficiently converted to the sulfonium conjugates **5** and **6** respectively. GTP-binding protein fragment Gα, which contains a free sulfhydryl group in an N-terminal cysteine residue, underwent smooth methionine labelling to **7** with concomitant oxidative disulfide formation. Teriparatide, a polypeptide containing two methionines, formed bis-sulfonium conjugate (**8**) with good conversions. α-Lactalbumin, a globular protein with a readily oxidized methionine residue,

undergoes bioconjugation to **9** with minimal competitive oxidation. Aprotinin (Trasylol) also forms the corresponding sulfonium conjugate **10** in good conversion to product. Particularly notable is the observation that two cysteine disulfide linkages within the structure of aprotinin (and the labelling to **7**) are not affected by **1b**, with a single methionine-derived conjugate obtained in high conversion. When the methionine residues are buried within the tertiary structure of the protein, for example with RNA-ase B, the bioconjugation does not occur, highlighting the inherent selectivity of the labelling process for exposed rather than inaccessible thioether groups. This feature is highlighted by the case of ubiquitin: the N-terminal methionine residue has only moderate surface exposure and can slow down the rate of functionalization to the point where oxidation becomes competitive. Accordingly, reaction under deoxygenated conditions enabled efficient conversion to the labelled product (84%), with a 10:1 label:oxidation ratio (**11**). The methionine-selective bioconjugation strategy effectively functionalizes a range of polypeptides and proteins in high conversion at micromolar concentrations and in very short reaction times. Given that **1b** is a carbon electrophile with two of the best leaving groups known to organic chemists, it would be thought to be very reactive; it is therefore remarkable that this species selectively engages the moderately nucleophilic methionine residue in the presence of competitively nucleophilic and oxidizable amino acid residues. Bioconjugation strategies that target other amino acids should, therefore, be compatible with, and complementary to, our methionine-functionalization process. To exemplify this, we showed that the GTP-binding protein fragment Gα could be first labelled at cysteine, using a maleimide derivative[23], and then conjugated at methionine using iodonium salt **1b** to form **12**. The methionine-selective process does not interfere with the cysteine–maleimide motif, which itself contains a thioether linkage, thereby highlighting possible applications towards multi-site heterolabelling of proteins[24].

The modular synthesis of the hypervalent iodine reagents enables facile incorporation of different acyl groups attached to the diazo motif, allowing the transfer of a range of functional payloads to proteins (Fig. 3b). Functional groups relevant to other bioorthogonal reactions are readily tolerated in both the reagent synthesis and the methionine labelling, smoothly affording sulfonium conjugates **4b** and **4c** with 95% and 85% conversion to product, respectively. Biochemical reporter groups, such as myristyl- and fluorescein-derived esters **4d** and **4e**, are also readily transferred to exenatide. Notably, we found that sulfonium conjugates of exenatide (such as **4f**) underwent reaction with the tertiary phosphine TCEP (tris(2-carboxyethyl)phosphine), a standard biochemical reagent, resulting in the cleavage of the labelling group and return of the parent exenatide **2b** in >90% conversion[25] (Fig. 3c); the cleavage reaction also works for conjugates **5**, **6** and **8** in comparable conversions and provides a stimulus-responsive means of removing the methionine label.

Next, we turned our attention to exploring the multifaceted reactivity that we anticipated would be intrinsic to the high-energy methionine-derived conjugate. The electrophilicity of the diazo sulfonium conjugate **4** prompted us to investigate the single-electron transfer chemistry of this reactive motif enabled by visible-light photocatalysis[26,27]. The addition of a single electron to the diazo sulfonium conjugate **4** could result in intermediate **15**, which, upon cleavage of the C–N_2 bond, would form a putative radical ylide synthon **15′** (Fig. 4a). We visualized two pathways through which we could exploit the reactivity of the previously unexplored radical ylide. First, combining the radical ylide with Hantzsch ester **13*** (from **13**) may lead to a reduction process resulting in the generation of a trialkylsulfonium motif, which would impart enhanced chemical stability to the protein conjugate. Furthermore, the use of a C-4 benzylated Hantzsch ester derivative (**14**, an established precursor for a benzyl radical)[28] to intercept the radical ylide species would lead to a C-benzylation product that could be used to introduce functional diversity to the protein conjugate. We screened a range of photocatalysts under visible-light conditions. When **3c** was irradiated with a 30 W lamp in the presence of *fac*-Ir(ppy)$_3$ (ppy, 2-phenylpyridinato) and the Hantzsch ester **13**[29,30], we observed the formation of the reduced trialkylsulfonium product **17a** in 95% yield (determined by ^{1}H NMR, Fig. 4b). Using these conditions, we showed that a range of sulfonium–protein conjugates, including exenatide, glucagon and thioredoxin derivatives, are reduced to stable trialkylsulfonium species with excellent conversions (**17b**–**e**, see Supplementary Information). Notably, reduction of a thioredoxin derivative[31] to its trialkylsulfonium–protein congener **17e** proceeds in high conversion without affecting its labile disulfide linkage, which serves to highlight the mild nature of this protocol. Additionally, the methionine bioconjugation and photoreduction steps can be carried out in a one-pot operation, which considerably simplifies the overall process without compromising the yield or purity of the trialkylsulfonium product.

In testing the viability of the proposed C-benzylation using a modified Hantzsch ester derivative[28], we found that treatment of the exenatide–sulfonium conjugate **4b** with *o*-tolyl Hantzsch ester derivative **14a**, under slightly modified photocatalytic conditions, led to cross-coupling and the formation of the C-ligation product **18a** in high conversion. This unique bioorthogonal protein functionalization reaction not only represents a synthetic radical–radical cross-coupling using a polypeptide scaffold, but also provides a platform for bioorthogonal protein diversification wherein two distinct functionalities could be introduced sequentially at the same amino acid residue.

In summary, through the merger of methionine-selective bioconjugation and a new visible-light-mediated photocatalytic reaction platform, information-rich synthetic constructs can be rapidly assembled by a two-step protocol directly from native proteins. The reactivity inherent to the methionine conjugate distinguishes the bioconjugation process from other methods. Moreover, the capacity for functional diversification, by tailoring the hypervalent iodine reagent and modified Hantzsch-ester derivative, means that highly functional protein conjugates could be made readily available directly from native proteins.

Data availability

The data that support the findings of this study are available within the paper and its Supplementary Information. Raw data are available from the corresponding author on reasonable request.

Received: 27 February 2017; Accepted: 21 August 2018;
Published online 15 October 2018.

1. Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J. Jr. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angew. Chem. Int. Ed*. **44**, 7342–7372 (2005).
2. Sletten, E. M. & Bertozzi, C. R. Bioorthogonal chemistry: fishing for selectivity in a sea of functionality. *Angew. Chem. Int. Ed*. **48**, 6974–6998 (2009).
3. Spicer, C. D. & Davis, B. G. Selective chemical protein modification. *Nat. Commun*. **5**, 4740 (2014).
4. Koniev, O. & Wagner, A. Developments and recent advancements in the field of endogenous amino acid selective bond forming reactions for bioconjugation. *Chem. Soc. Rev*. **44**, 5495–5551 (2015).
5. Dawson, P. E. & Kent, S. B. H. Synthesis of native proteins by chemical ligation. *Annu. Rev. Biochem*. **69**, 923–960 (2000).
6. Lang, K. & Chin, J. W. Cellular incorporation of unnatural amino acids and bioorthogonal labeling of proteins. *Chem. Rev*. **114**, 4764–4806 (2014).
7. Wang, L., Xie, J. & Schultz, P. G. Expanding the genetic code. *Annu. Rev. Biophys. Biomol. Struct*. **35**, 225–249 (2006).
8. Vinogradova, E. V., Zhang, C., Spokoyny, A. M., Pentelute, B. L. & Buchwald, S. L. Organometallic palladium reagents for cysteine bioconjugation. *Nature* **526**, 687–691 (2015).
9. Wright, T. H. et al. Posttranslational mutagenesis: a chemical strategy for exploring protein side-chain diversity. *Science* **354**, aag1465 (2016).
10. Yang, A. et al. A chemical biology route to site-specific authentic protein modifications. *Science* **354**, 623–626 (2016).
11. Abegg, D. et al. Proteome-wide profiling of targets of cysteine reactive small molecules by using ethynyl benziodoxolone reagents. *Angew. Chem. Int. Ed*. **54**, 10852–10857 (2015).
12. Levine, R. L., Moskovitz, J. & Stadtman, E. R. Oxidation of methionine in proteins: roles in antioxidant defense and cellular regulation. *IUBMB Life* **50**, 301–307 (2000).

13. Cowie, D. B., Cohen, G. N., Bolton, E. T. & De Robichon-Szulmajster, H. Amino acid analog incorporation into bacterial proteins. *Biochim. Biophys. Acta* **34**, 39–46 (1959).
14. Lin, S. et al. Redox-based reagents for chemoselective methionine bioconjugation. *Science* **355**, 597–602 (2017).
15. Gross, E. & Witkop, B. Nonenzymatic cleavage of peptide bonds: the methionine residues in bovine pancreatic ribonuclease. *J. Biol. Chem*. **237**, 1856–1860 (1962).
16. Gundlach, H. G., Stein, W. H. & Moore, S. The nature of the amino acid residues involved in the inactivation of ribonuclease by iodoacetate. *J. Biol. Chem*. **234**, 1754–1760 (1959).
17. Vithayathil, P. J. & Richards, F. M. Modification of the methionine residue in the peptide component of ribonuclease-S. *J. Biol. Chem*. **235**, 2343–2351 (1960).
18. Kramer, J. R. & Deming, T. J. Preparation of multifunctional and multireactive polypeptides via methionine alkylation. *Biomacromolecules* **13**, 1719–1723 (2012).
19. Kramer, J. R. & Deming, T. J. Reversible chemoselective tagging and functionalization of methionine containing peptides. *Chem. Commun*. **49**, 5144–5146 (2013).
20. Stang, P. J. & Zhdankin, V. V. Organic polyvalent iodine compounds. *Chem. Rev.* **96**, 1123–1178 (1996).
21. Weiss, R., Seubert, J. & Hampel, F. α-Aryliodonio diazo compounds: S_N reactions at the α-C atom as a novel reaction type for diazo compounds. *Angew. Chem. Int. Edn Engl*. **33**, 1952–1953 (1994).
22. Schnaars, C., Hennum, M. & Bonge-Hansen, T. Nucleophilic halogenations of diazo compounds, a complementary principle for the synthesis of halodiazo compounds: experimental and theoretical studies. *J. Org. Chem*. **78**, 7488–7497 (2013).
23. Kim, Y. et al. Efficient site-specific labeling of proteins via cysteines. *Bioconjug. Chem*. **19**, 786–791 (2008).
24. Mülhberg, M. et al. Orthogonal dual-modification of proteins for the engineering of multivalent protein scaffolds. *Beilstein J. Org. Chem*. **11**, 784–791 (2015).
25. Staudinger, H. & Lüscher, G. Über darstellung und reaktionen von phosphazinen. *Helv. Chim. Acta* **5**, 75–86 (1922).
26. Prier, C. K., Rankic, D. A. & MacMillan, D. W. C. Visible light photoredox catalysis with transition metal complexes: applications in organic synthesis. *Chem. Rev.* **113**, 5322–5363 (2013).
27. Chen, Y., Kamlet, A. S., Steinman, J. B. & Liu, D. R. A biomolecule-compatible visible-light-induced azide reduction from a DNA-encoded reaction-discovery system. *Nat. Chem*. **3**, 146–153 (2011).
28. Huang, W. & Cheng, X. Hantzsch esters as multifunctional reagents in visible-light photoredox catalysis. *Synlett* **28**, 148–158 (2017).
29. Fukuzumi, S., Hironaka, K. & Tanaka, T. Photoreduction of alkyl halides by an NADH model compound. An electron-transfer chain mechanism. *J. Am. Chem. Soc*. **105**, 4722–4727 (1983).
30. Hedstrand, D. M., Kruizinga, W. H. & Kellog, R. M. Light induced and dye accelerated reductions of phenacyl onium salts by 1,4-dihydropyridines. *Tetrahedron Lett*. **19**, 1255–1258 (1978).
31. Krause, G., Lundström, J., Barea, J. L., Pueyo de la Cuesta, C. & Holmgren, A. Mimicking the active site of protein disulfide-isomerase by substitution of proline 34 in *Escherichia coli* thioredoxin. *J. Biol. Chem*. **266**, 9494–9500 (1991).

Acknowledgements We thank M. Nappi and C. Guerot for advice and useful discussions. We thank the Marie Curie Actions program (M.T.T. and M.G.S.), AstraZeneca and EPRSC (J.E.N.), and the European Research Council (ERC-SRG-259711), EPSRC (EP/100548X/1) and the Royal Society (Wolfson Merit Award) for fellowships (M.J.G.). We are grateful to J. Chin, N. Huguen, M. Skehel, H. Lewis and M. Edgeworth for assistance with protein purification and mass spectrometry experiments.

Reviewer information *Nature* thanks A. Spokoyny and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.J.G., M.T.T., J.E.N. and M.G.S. conceived the project and designed the experiments. M.J.G., M.T.T., J.E.N. and M.G.S. performed and analysed the experiments. M.J.G., M.T.T. and J.E.N. wrote the paper.

Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0608-y.
Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.J.G.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

LETTER

https://doi.org/10.1038/s41586-018-0614-0

CO_2 storage and release in the deep Southern Ocean on millennial to centennial timescales

J. W. B. Rae[1]*, A. Burke[1], L. F. Robinson[2], J. F. Adkins[3], T. Chen[2,4], C. Cole[1], R. Greenop[1], T. Li[2,4], E. F. M. Littley[1], D. C. Nita[1,5], J. A. Stewart[1,2] & B. J. Taylor[1]

The cause of changes in atmospheric carbon dioxide (CO_2) during the recent ice ages is yet to be fully explained. Most mechanisms for glacial–interglacial CO_2 change have centred on carbon exchange with the deep ocean, owing to its large size and relatively rapid exchange with the atmosphere[1]. The Southern Ocean is thought to have a key role in this exchange, as much of the deep ocean is ventilated to the atmosphere in this region[2]. However, it is difficult to reconstruct changes in deep Southern Ocean carbon storage, so few direct tests of this hypothesis have been carried out. Here we present deep-sea coral boron isotope data that track the pH—and thus the CO_2 chemistry—of the deep Southern Ocean over the past forty thousand years. At sites closest to the Antarctic continental margin, and most influenced by the deep southern waters that form the ocean's lower overturning cell, we find a close relationship between ocean pH and atmospheric CO_2: during intervals of low CO_2, ocean pH is low, reflecting enhanced ocean carbon storage; and during intervals of rising CO_2, ocean pH rises, reflecting loss of carbon from the ocean to the atmosphere. Correspondingly, at shallower sites we find rapid (millennial- to centennial-scale) decreases in pH during abrupt increases in CO_2, reflecting the rapid transfer of carbon from the deep ocean to the upper ocean and atmosphere. Our findings confirm the importance of the deep Southern Ocean in ice-age CO_2 change, and show that deep-ocean CO_2 release can occur as a dynamic feedback to rapid climate change on centennial timescales.

The Southern Ocean may act as a net source of CO_2 from the deep ocean to the atmosphere or as a net sink[3], depending on the balance between regional CO_2 supply via circulation and CO_2 removal via biological productivity. Various records have shown that large changes in circulation[4,5] and biological productivity[6] occurred in the Southern Ocean on glacial timescales, with the potential to change the partitioning of carbon between the deep ocean and the atmosphere. However, reconstructions of deep ocean CO_2 chemistry are currently sparse and harder to interpret simply in terms of carbon storage. For instance, records of CO_2 chemistry from the deep Atlantic[7] and deep Pacific[8] show decreases in carbonate ion saturation and pH during millennial-scale intervals of atmospheric CO_2 rise; in the absence of other processes, low carbonate ion concentration and pH imply an increase in CO_2 storage in the deep ocean, so these signals are thought instead to be dominated by changes in circulation and deep water masses. On longer timescales, records from the deep Indo-Pacific[9] appear to reflect changes in CO_2 storage, but are damped by the buffering influence of carbonate compensation. Records of deep ocean CO_2 chemistry that clearly demonstrate CO_2 storage during falls in atmospheric CO_2, and CO_2 release during increases in atmospheric CO_2, have proved elusive.

Here we test the hypothesis that carbon storage in the deep Southern Ocean played an important role in ice-age CO_2 change, with new boron isotope ($\delta^{11}B$) data from uranium–thorium-dated deep-sea corals from the Drake Passage[4] (Fig. 1; see Methods). The boron isotope pH proxy (see Methods) provides a sensitive measure of the ocean carbonate

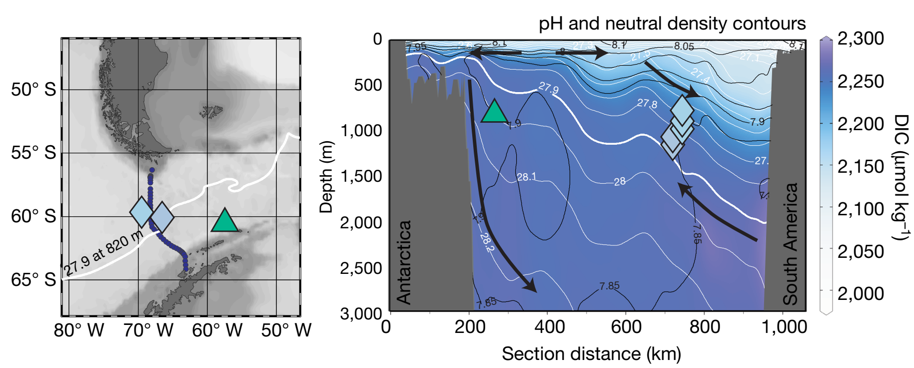


Fig. 1 | Locations of deep-sea coral samples. The cross-section (right) was constructed from hydrographic stations across the Drake Passage (dark blue dots on map, left). Steeply dipping isopycnals in this region (white contours) mean that our sites fall into two groups that span distinct volumes of the deep ocean[11]. The green triangle marks the lower cell sites, which lie close to the Antarctic continental margin in the Shackleton Fracture Zone; the blue diamonds mark the upper cell sites, which lie at lower densities on the Sars (lighter blue) and Interim (darker blue) seamounts. Lower cell waters are rich in DIC (shading, right) with low pH (black contours); upper cell waters have higher pH and are more closely connected to the atmosphere. Coral locations on the section are given in coordinates of depth and neutral density, based on water column data collected alongside the coral dredges. Note that there is no significant offset in our $\delta^{11}B$ data between upper cell corals from different depths (Extended Data Fig. 2), and that differences in pH between these sites are small compared to the range seen in our records. The 27.9 neutral density contour, which today lies at the boundary between the lower and upper cells[11], is shown in the map view (left) at 820 m, the average depth of our corals. Section data are from GLODAPv2[34,35], plotted using isopycnal gridding in Ocean Data View.

[1]School of Earth and Environmental Sciences, University of St Andrews, St Andrews, UK. [2]School of Earth Sciences, University of Bristol, Bristol, UK. [3]Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA, USA. [4]School of Earth Sciences and Engineering, Nanjing University, Nanjing, China. [5]Faculty of Environmental Science and Engineering, Babes-Bolyai University, Cluj-Napoca, Romania. *e-mail: jwbr@st-andrews.ac.uk

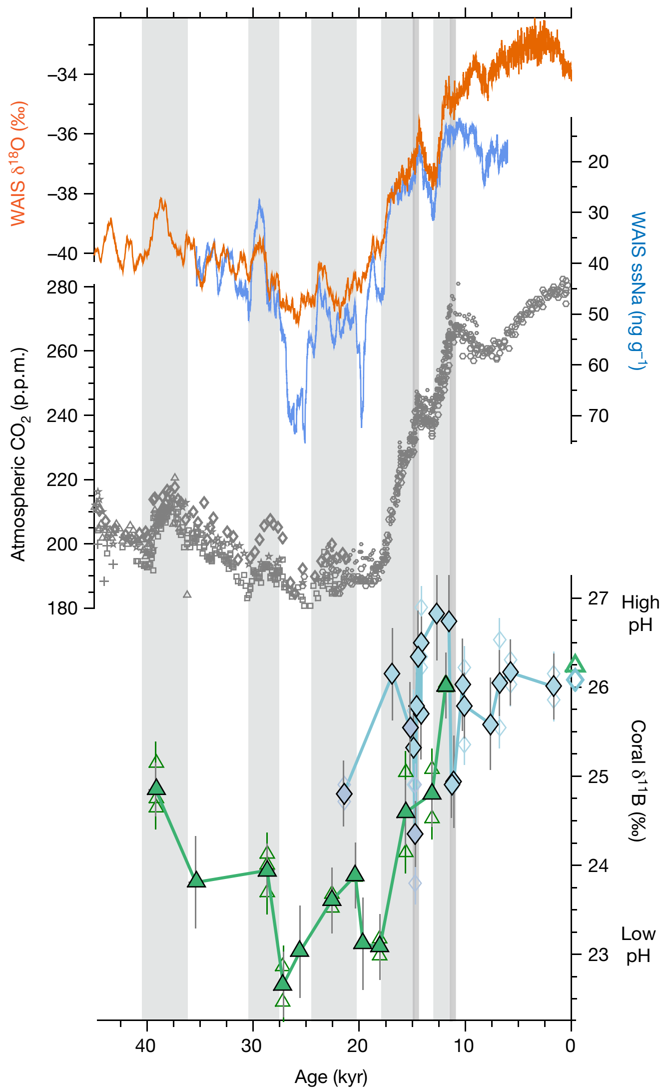


Fig. 2 | Deep Southern Ocean CO_2 chemistry, atmospheric CO_2, and Antarctic climate records over the past 40,000 years. Green triangles and blue diamonds show lower and upper cell deep-sea coral $\delta^{11}B$ data, respectively. Individual subsamples are shown as small open symbols and mean values as larger filled symbols. Error bars on individual subsamples are equivalent to 2 s.d. analytical reproducibility and error bars on mean coral values represent 2 s.e. uncertainty on the mean of replicate subsamples (see Methods). Synchronized ice core CO_2 data[36] are shown in grey symbols: circles from Dome C, dots from WAIS, stars from Taylor Dome, triangles from TALDICE, pluses from EDML, diamonds from Byrd, and squares from Siple Dome. West Atlantic ice sheet (WAIS) $\delta^{18}O$ (orange line), which reflects Antarctic temperature, and sea salt sodium (ssNa, blue line), a proxy for sea ice, have been smoothed with a running mean[28]. Grey bands highlight intervals of CO_2 rise.

system, closely tracking CO_2 concentrations and reflecting the ratio of the two master variables, dissolved inorganic carbon (DIC) and alkalinity. Although full reconstruction of the carbonate system requires knowledge of a second parameter, it is unlikely that alkalinity was lower in the glacial ocean[10], or varied as dynamically as DIC, so our $\delta^{11}B$ pH record can be largely attributed to changes in carbon storage. Note that as our $\delta^{11}B$ record extends beyond the pH calibration possible in modern *Desmophyllum dianthus* (see Methods), we focus our discussion on relative changes in pH as traced by coral $\delta^{11}B$, and provide absolute pH estimates in Extended Data Fig. 1 for reference. Our sample sites reflect distinct volumes of the deep ocean[11]: the 'lower cell' sites lie close to the Antarctic continental margin, bathed by waters that plumb the middle to lower depths of the deep ocean; the 'upper cell' sites lie on lighter isopycnal surfaces, bathed by waters found at shallower depths in the ocean basins (Fig. 1 and Extended Data Fig. 2). Note that there is no significant offset in our $\delta^{11}B$ data between upper cell corals from different depths (Extended Data Fig. 2), and that differences in pH between these sites are small (Fig. 1) compared to the range seen in our records.

During the last glacial maximum (LGM) and early deglaciation, we see a clear gradient between the lower cell sites, which show low $\delta^{11}B$ and pH, and the upper cell sites, with relatively high $\delta^{11}B$ and pH (Fig. 2). This supports the idea that during glacial intervals the deep ocean—and its carbon—was more stratified into two cells with limited interactions[11]. Our data show that the lower cell was rich in carbon compared to the upper cell and compared to modern values expected at this site (Fig. 1, Extended Data Fig. 1), providing strong support for the hypothesis that the deep glacial ocean sequestered carbon from the upper ocean and the atmosphere[2].

During the deglaciation, this gradient in deep carbon breaks down, with lower cell pH rising in step with atmospheric CO_2, and pH in the upper cell falling towards lower cell values (Fig. 3). This provides direct evidence for the transfer of carbon from the deep ocean to the upper ocean and the atmosphere. Carbon transfer to the upper ocean appears to be particularly pronounced at approximately 14.7 and 11.7 thousand years ago (kyr ago), coincident with the centennial-scale jumps in atmospheric CO_2[12] associated with abrupt warming in the Northern Hemisphere. This provides the first evidence, to our knowledge, of a fast teleconnection between abrupt changes in the North Atlantic and the carbon chemistry of the deep Southern Ocean. Lower cell *D. dianthus* samples have not been found in the Holocene (see Methods), but the available data at the end of the deglaciation and in the modern water column suggest much weaker pH gradients, consistent with less-pronounced property gradients in the modern deep ocean compared to the glacial deep ocean[11,13,14].

Our data show that the carbonate chemistry of the deep Southern Ocean was closely linked to atmospheric CO_2 changes over the past 40,000 years. These data thus provide a crucial missing piece of the glacial CO_2 puzzle: the most direct evidence to date of deep Southern Ocean carbon storage and release, as previously inferred from physical properties[5,13], carbon isotopes[15], and oxygen content[16] (Fig. 2, Extended Data Fig. 3). While other processes[3] and regions[8] may have contributed to the full magnitude of glacial–interglacial CO_2 change, our data demonstrate a key role for the Southern Ocean on millennial to centennial timescales.

Several processes might have contributed to the changes in CO_2 storage observed in our record, including changes in ventilation[4,16], biological pump efficiency[6], and sea ice[17] (Extended Data Fig. 3). We note a close correspondence between lower cell pH and ice core sea salt sodium, a proxy that may reflect changes in sea ice production[18,19] (Fig. 2), which suggests that sea ice may have played an important role in CO_2 change. Sea ice has the potential to influence CO_2 storage both through its influence as a 'lid' on surface-ocean outgassing[17], and its effect on deep circulation[11]. Expansion of sea ice at the LGM[11], alongside an increase in surface ocean density in the Southern Ocean relative to the North Atlantic[20], would have helped to create an expanded lower cell with salty[13], CO_2-rich water. This could have helped to shoal the upper–lower cell boundary above the zone of enhanced mixing over rough bottom topography[11,14], trapping salt and CO_2 in the abyss. Accumulation of CO_2 at depth would have been further promoted by an enhanced biological pump due to iron fertilization[6] and increased upper ocean stratification, which would also have reduced CO_2 escape through leads and under ice-free conditions.

This framework may also explain the release of CO_2 from the deep Southern Ocean on millennial timescales[21]. Increases in CO_2 typically occur during intervals of cold stadial conditions in the Northern Hemisphere and warming in the south (the bipolar seesaw)[22]. This southern warming is associated with a decrease in southern sea ice (Figs. 2, 3) and a decrease in the surface density gradient between the Southern Ocean and the North Atlantic, shifting the boundary between the overturning cells to greater depth in the basins[11,20]. CO_2-rich water that was previously isolated in the abyss could thereby have been mixed

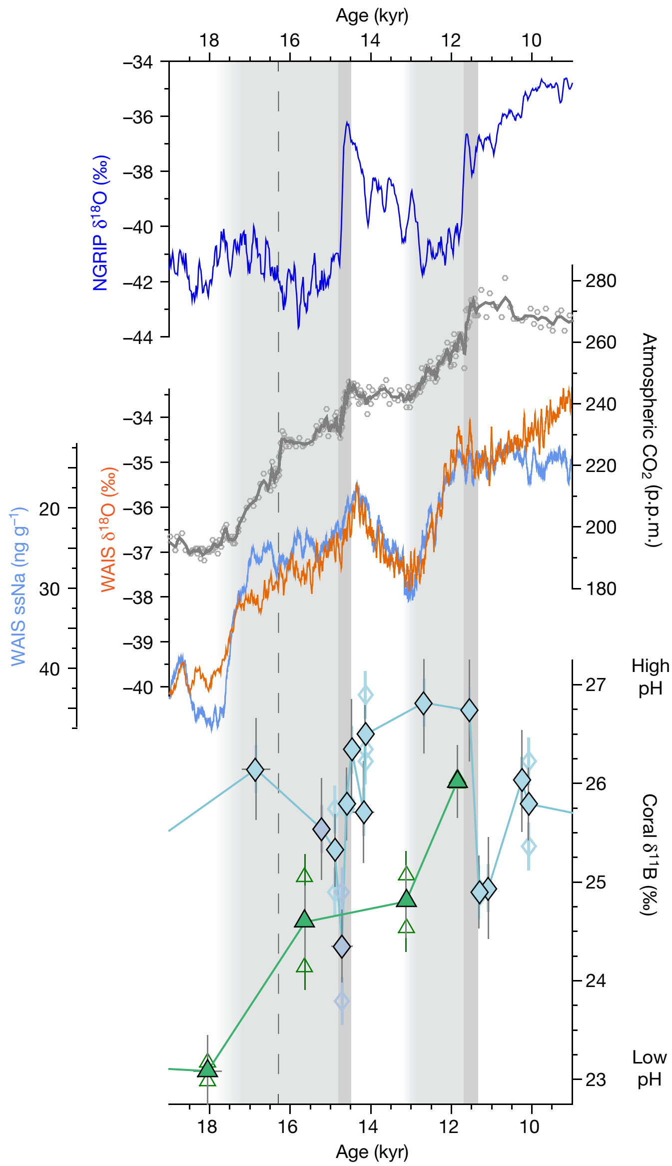


Fig. 3 | Deglacial records of deep Southern Ocean CO_2 chemistry, atmospheric CO_2, and climate over Antarctica and Greenland. Green triangles and blue diamonds show lower and upper cell deep-sea coral $\delta^{11}B$ data, respectively. Individual subsamples are shown as small open symbols and mean values as larger filled symbols. Error bars on individual subsamples are equivalent to 2 s.d. analytical reproducibility and error bars on mean coral values represent 2 s.e. uncertainty on the mean of replicate subsamples (see Methods). Greenland ice core $\delta^{18}O$ (dark blue line), WAIS $\delta^{18}O$ (orange line), and sea salt sodium (blue line) have been smoothed with a running mean[28]. CO_2 data (grey symbols) are from WAIS[12] with a five-point running mean (grey line). Light grey vertical bands highlight intervals of millennial-scale CO_2 rise; dark grey vertical bands highlight intervals of centennial-scale CO_2 rise associated with North Hemisphere warming; and the vertical dashed line indicates the rapid CO_2 rise event that occurred at 16.3 kyr ago, within Heinrich Stadial 1.

into the upper cell over rough topography in the ocean basins, and/or transferred into the upper cell upon upwelling north of the sea ice edge, perhaps aided by the westerly winds[23,24] or increased mixed layer depths in the Southern Ocean. CO_2 loss from the deep ocean might also have been aided by reduced biological pump efficiency[6]. Whatever the exact mechanism, this carbon transfer is recorded by a pH increase in our lower cell corals and a pH decrease in our upper cell corals, as CO_2 was transferred to the upper ocean and the atmosphere. Note that as a southward shift in the fronts at these times might be expected to expose our sites to higher-pH water (Fig. 1), the transfer of low-pH DIC-rich water into the upper cell might have been even larger than observed. Indeed, a southward frontal shift and breakdown in stratification are likely to explain the younging seen in upper cell radiocarbon[4,25] at this time (Extended Data Fig. 4). Upwelling of carbon and nutrient-rich water during cold Northern Hemisphere stadials is also supported by low pH in surface waters, as recorded by boron isotopes in planktic foraminifera[26] and enhanced opal fluxes[23] (Fig. 3, Extended Data Fig. 4). If salt from the high-salinity lower cell[13] was also transferred back into the upper cell, this might have aided the re-initiation of North Atlantic deep water (NADW) formation[27]. Once interstadial conditions were reestablished in the North Atlantic, the Southern Ocean would have started to cool via the bipolar seesaw and sea ice expanded[28]. This would shoal the cell boundary, reduce ocean–atmosphere exchange, and allow CO_2 and salt to again become trapped in the abyss (Fig. 4).

The centennial-scale increases in CO_2 at 14.7 and 11.7 kyr ago[12], which are associated with pronounced minima in upper cell pH (Fig. 3), appear to require a more efficient mode of Southern Ocean CO_2 release associated with abrupt Northern Hemisphere warming[25]. High-resolution records of Antarctic deuterium excess indicate that there was a northward shift in the southern westerly winds synchronous with abrupt Northern Hemisphere warming[29]; by contrast, the bipolar seesaw cooling response in the south is lagged, with Antarctic $\delta^{18}O$ and sea salt sodium taking around 200 years to show the onset of cooling conditions and increased sea ice production[28] (Fig. 3). This might have led to a transient condition in the Southern Ocean where CO_2 could be efficiently mixed up from the lower cell[24], owing to the fast shift in the winds, and then outgassing unimpeded by sea ice, which had not yet expanded[17]. The northward shift in the fronts might also have contributed to the pH minima by exposing our upper cell sites to lower-pH water from the south (Fig. 1), but such a shift would have to have been large given that upper cell pH appears to have reached values similar to, or even lower than, the lower cell at this time, suggesting that increased input of CO_2-rich water was required. Increased input of previously isolated carbon-rich waters is also seen in radiocarbon data[4,25], which show an interruption of their deglacial younging and a slight increase in age during these events (Extended Data Fig. 4). Whatever the exact mechanisms involved, our data demonstrate that the Southern Ocean can fill with CO_2-rich waters on centennial timescales and can therefore give out its carbon rapidly, countering suggestions that centennial-scale increases in CO_2 are too quick for a deep ocean driver and require exogenous carbon addition (such as from methane hydrates or the terrestrial biosphere[30]). Our data show that rapid changes in the Southern Ocean acted in concert with resumption of vigorous Atlantic meridional overturning circulation (AMOC)[25] to drive rapid increases in CO_2.

Although our lower cell $\delta^{11}B$ pH data generally show a close coupling with atmospheric CO_2, this relationship was muted during peak glacial conditions, with pronounced minima in pH at approximately 26 and 20 kyr ago, alongside extensive sea ice[28], low upwelling[23], and an efficient biological pump[6] (Extended Data Fig. 3), but minimal change in atmospheric CO_2. This supports the idea that there is a lower limit on atmospheric CO_2 at about 190 p.p.m.[31]; although Southern Ocean carbon storage continues to increase, its influence on the atmosphere appears to be offset by other processes, perhaps the onset of CO_2-limitation of primary productivity on land[31].

Overall, our data provide a clear demonstration that storage and release of CO_2 in the deep Southern Ocean played a central role in glacial–interglacial changes in atmospheric CO_2. These changes in ocean CO_2 storage are likely to have been driven by a combination of changes in ocean circulation, biological pump efficiency, and sea ice cover. We note a close correspondence between CO_2 storage and ice core sea ice sodium records, which may suggest that Southern Ocean sea ice played a key role in glacial CO_2 change, owing to its joint influence on deep overturning and surface outgassing. This provides a mechanistic explanation for the tight link between Antarctic temperature and CO_2 change on glacial–interglacial timescales, although several processes acting together are likely to be required to explain the full magnitude of glacial CO_2 change. Our data also

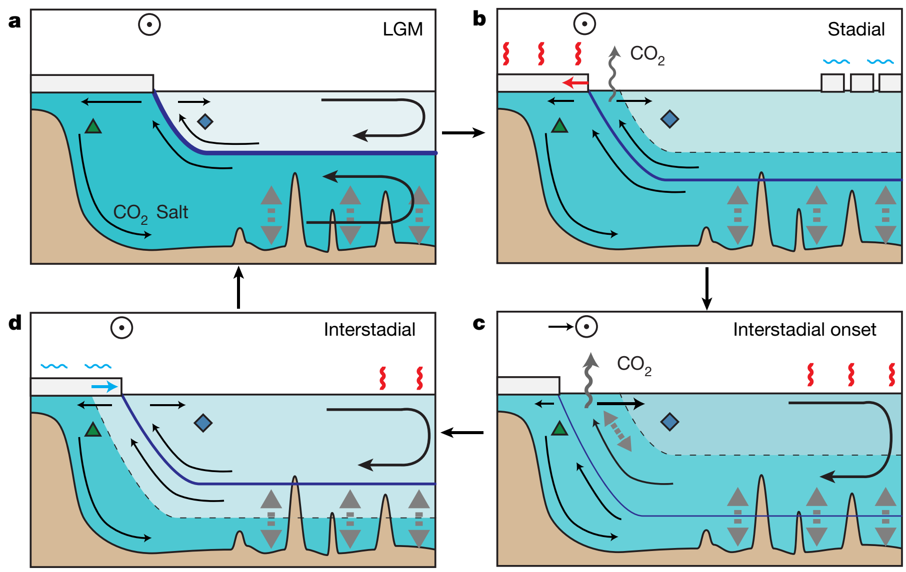


Fig. 4 | Schematic of changes in sea ice, circulation, and deep ocean carbon storage. a, At the LGM, Southern Ocean cooling and extensive sea ice helped to create an expanded lower cell with salty, CO_2-rich water. The cell boundary was shoaled above the zone of enhanced mixing over rough bottom topography, isolating salt and CO_2 in the abyss. **b**, During Northern Hemisphere stadial events, North Atlantic overturning was reduced, the Southern Ocean warmed, and sea ice retreated. This shifted the cell boundary such that water previously isolated in the lower cell now upwells north of the sea ice edge. This water, and its CO_2 and salt, were transferred to the upper ocean and CO_2 outgassed to the atmosphere. Transfer of salt from the lower to the upper cell might have helped to re-initiate NADW formation. **c**, At the onset of a Northern Hemisphere interstadial event (for example, Bølling–Allerød, end Younger Dryas), resumption of NADW warmed the Northern Hemisphere and led to a rapid northward shift in the westerly winds; the Southern Ocean temperature and sea ice response was slower. This might have created a transient condition in which sea ice was unable to shield the ocean from enhanced isopycnal mixing nor the atmosphere from enhanced outgassing, leading to a centennial-scale increase in CO_2. **d**, As the interstadial continued, the Southern Ocean cooled and sea ice expanded. This shoaled the cell boundary and allowed salt and carbon to again become trapped in the abyss. Climate states with moderate sea ice extent, in which the cell boundary hovers around the top of rough seafloor topography, may give favourable conditions for rapid climate and CO_2 change, as the ocean flips between modes of connection and isolation of the upper and lower cells.

highlight the ability of the Southern Ocean—and its CO_2—to respond to millennial and centennial-scale shifts in climate linked to the North Atlantic's overturning circulation. Indeed, it is possible that the framework presented here, which links the storage and release of deep ocean CO_2 and salt to changes in Southern Ocean sea ice and the bipolar seesaw, may help to account for the occurrence of millennial-scale changes in CO_2 and climate during mid-glacial conditions. At these times, moderate sea ice extent means that the boundary between the ocean's lower and upper cells was located near the top of rough seafloor topography. Small shifts in the cell boundary, linked to changes in the Southern sea ice edge, could therefore have driven large shifts in the degree of mixing between the lower and upper cells, and their salt and CO_2. This framework may also give behaviour similar to 'density oscillator' models for rapid climate change[27,32,33]: when the AMOC is active and the Northern Hemisphere warms, the south cools and sea ice expands, progressively isolating cold salty water (and CO_2) in the lower cell. This may make the AMOC vulnerable to collapse, at which point the south warms, sea ice retreats, and the cell boundary deepens. This helps to mix salt (and CO_2) back into the upper cell and makes the lower cell fresher and warmer, reducing the deep ocean density contrast, and helping to poise the system for AMOC resumption.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0614-0.

Received: 6 March 2018; Accepted: 29 August 2018;
Published online 24 October 2018.

1. Broecker, W. S. Glacial to interglacial changes in ocean chemistry. *Prog. Oceanogr.* **11**, 151–197 (1982).
2. Sarmiento, J. L. & Toggweiler, J. R. A new model for the role of the oceans in determining atmospheric pCO_2. *Nature* **308**, 621–624 (1984).
3. Sigman, D. M., Hain, M. P. & Haug, G. H. The polar ocean and glacial cycles in atmospheric CO_2 concentration. *Nature* **466**, 47–55 (2010).
4. Burke, A. & Robinson, L. F. The Southern Ocean's role in carbon exchange during the last deglaciation. *Science* **335**, 557–561 (2012).
5. Roberts, J. et al. Evolution of South Atlantic density and chemical stratification across the last deglaciation. *Proc. Natl Acad. Sci. USA* **113**, 514–519 (2016).
6. Martinez-Garcia, A. et al. Iron fertilization of the Subantarctic Ocean during the last ice age. *Science* **343**, 1347–1350 (2014).
7. Yu, J. et al. Deep South Atlantic carbonate chemistry and increased interocean deep water exchange during last deglaciation. *Quat. Sci. Rev.* **90**, 80–89 (2014).
8. Rae, J. W. B. et al. Deep water formation in the North Pacific and deglacial CO2 rise. *Paleoceanography* **29**, 645–667 (2014).
9. Yu, J. et al. Responses of the deep ocean carbonate system to carbon reorganization during the last glacial–interglacial cycle. *Quat. Sci. Rev.* **76**, 39–52 (2013).
10. Rickaby, R. E. M., Elderfield, H., Roberts, N., Hillenbrand, C. D. & Mackensen, A. Evidence for elevated alkalinity in the glacial Southern Ocean. *Paleoceanography* **25**, PA1209 (2010).
11. Ferrari, R. et al. Antarctic sea ice control on ocean circulation in present and glacial climates. *Proc. Natl Acad. Sci. USA* **111**, 8753–8758 (2014).
12. Marcott, S. A. et al. Centennial-scale changes in the global carbon cycle during the last deglaciation. *Nature* **514**, 616–619 (2014).
13. Adkins, J. F., McIntyre, K. & Schrag, D. P. The salinity, temperature, and $\delta^{18}O$ of the glacial deep ocean. *Science* **298**, 1769–1773 (2002).
14. Burke, A., Stewart, A. L., Adkins, J. F. & Ferrari, R. The glacial mid-depth radiocarbon bulge and its implications for the overturning circulation. *Paleoceanography* **30**, 1021–1039 (2015).
15. Charles, C. D. et al. Millennial scale evolution of the Southern Ocean chemical divide. *Quat. Sci. Rev.* **29**, 399–409 (2010).
16. Jaccard, S. L., Galbraith, E. D., Martinez-Garcia, A. & Anderson, R. F. Covariation of deep Southern Ocean oxygenation and atmospheric CO_2 through the last ice age. *Nature* **530**, 207–210 (2016).
17. Stephens, B. B. & Keeling, R. F. The influence of Antarctic sea ice on glacial-interglacial CO_2 variations. *Nature* **404**, 171–174 (2000).
18. Wolff, E. W. et al. Southern Ocean sea-ice extent, productivity and iron flux over the past eight glacial cycles. *Nature* **440**, 491–496 (2006).
19. Abram, N. J., Wolff, E. W. & Curran, M. A. J. A review of sea ice proxy information from polar ice cores. *Quat. Sci. Rev.* **79**, 168–183 (2013).
20. Galbraith, E. & de Lavergne, C. Response of a comprehensive climate model to a broad range of external forcings: relevance for deep ocean ventilation and the development of late Cenozoic ice ages. *Clim. Dyn.* https://doi.org/10.1007/s00382-018-4157-8 (2018).

21. Ahn, J. & Brook, E. J. Atmospheric CO_2 and climate on millennial time scales during the last glacial period. *Science* **322**, 83–85 (2008).
22. Stocker, T. F. The seesaw effect. *Science* **282**, 61–62 (1998).
23. Anderson, R. F. et al. Wind-driven upwelling in the Southern Ocean and the deglacial rise in atmospheric CO_2. *Science* **323**, 1443–1448 (2009).
24. Abernathey, R. & Ferreira, D. Southern Ocean isopycnal mixing and ventilation changes driven by winds. *Geophys. Res. Lett.* **42**, 10,357–10,365 (2015).
25. Chen, T. et al. Synchronous centennial abrupt events in the ocean and atmosphere during the last deglaciation. *Science* **349**, 1537–1541 (2015).
26. Martínez-Botí, M. A. et al. Boron isotope evidence for oceanic carbon dioxide leakage during the last deglaciation. *Nature* **518**, 219–222 (2015).
27. Broecker, W. S., Bond, G., Klas, M., Bonani, G. & Wolfli, W. A salt oscillator in the glacial Atlantic? 1. The concept. *Paleoceanography* **5**, 469–477 (1990).
28. Members, W. D. P. et al. Precise interpolar phasing of abrupt climate change during the last ice age. *Nature* **520**, 661–665 (2015).
29. Markle, B. R. et al. Global atmospheric teleconnections during Dansgaard-Oeschger events. *Nat. Geosci.* **10**, 36–40 (2017).
30. Köhler, P., Knorr, G. & Bard, E. Permafrost thawing as a possible source of abrupt carbon release at the onset of the Bølling/Allerød. *Nat. Commun.* **5**, 5520 (2014).
31. Galbraith, E. D. & Eggleston, S. A lower limit to atmospheric CO_2 concentrations over the past 800,000 years. *Nat. Geosci.* **10**, 295–298 (2017).
32. Bereiter, B., Shackleton, S., Baggenstos, D., Kawamura, K. & Severinghaus, J. Mean global ocean temperatures during the last glacial transition. *Nature* **553**, 39–44 (2018).
33. Keeling, R. F. & Stephens, B. B. Antarctic sea ice and the control of Pleistocene climate instability. *Paleoceanography* **16**, 112–131 (2001).
34. Key, R. M. et al. Global Ocean Data Analysis Project, Version 2 (GLODAPv2). http://doi.org/10.3334/CDIAC/OTG.NDP093_GLODAPv2 (2015).
35. Olsen, A. et al. The Global Ocean Data Analysis Project version 2 (GLODAPv2)—an internally consistent data product for the world ocean. *Earth Syst. Sci. Data* **8**, 297–323 (2016).
36. Bereiter, B. et al. Revision of the EPICA Dome C CO_2 record from 800 to 600 kyr before present. *Geophys. Res. Lett.* **42**, 542–549 (2015).

Acknowledgements This work was supported by NERC Standard Grant NE/N003861/1 to J.W.B.R. and L.F.R., an NOAA Climate and Global Change VSP Fellowship to J.W.B.R, NERC Standard Grant NE/M004619/1 to A.B. and J.W.B.R., a NERC Strategic Environmental Science Capital Grant to A.B. and J.W.B.R., Marie Curie Career Integration Grant CIG14-631752 to A.B., an ERC consolidator grant to L.F.R., NSF grant OCE-1503129 to J.F.A., and NERC studentships to B.T. and E.L.

Reviewer information *Nature* thanks C. Buizert and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.W.B.R., A.B., and L.F.R. designed the study. A.B., L.F.R., T.C., and T.L. collected and uranium–thorium dated the coral samples. J.W.B.R., B.T., E.L., C.C., R.G., J.A.S., and D.C.N. made boron isotope analyses. J.W.B.R., A.B., L.F.R., and J.F.A. developed the interpretation and all authors contributed to the preparation of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0614-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0614-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.W.B.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Sample collection and chronology. Deep-sea coral samples were collected by dredge during two cruises in the Drake Passage (NBP0805 and NBP1103). Sample locations are given in Extended Data Tables 1, 2, and sample depths are also shown in Extended Data Fig. 2. A total of 392 *D. dianthus* samples were initially 'reconnaissance' dated, either by ^{14}C[37,38] or laser ablation U-Th[25,39], to obtain preliminary ages. Suitable samples with ages within the past ~50 kyr were then precisely dated using isotope dilution U–Th by multicollector-inductively coupled plasma mass spectrometry (MC-ICPMS)[4,25]. All ages have been published previously, and have typical uncertainties of around ± 1% (2 s.d.), although this varies between samples depending on their initial ^{230}Th (calculated from measured ^{232}Th and assuming an initial atomic $^{232}Th/^{230}Th$ ratio of 12,500 ± 12,500). Age errors are plotted in all time series figures but are typically smaller than the symbols. This precisely dated and unbioturbated deep sea archive provides a unique record of ocean pH change at resolution comparable to those of the ice cores.

Sample preparation. Coral pieces were sampled from the growing tip of coral septa. Samples were physically cleaned using a Dremmel tool to remove all visible iron-manganese oxides and any chalky white carbonate, indicative of alteration.

We tested the potential influence of microstructural variability on coral $\delta^{11}B$ with multiple solid sub-samples from the same coral (Extended Data Fig. 6). Coral centres of calcification have previously been observed to have anomalously light boron isotope values, along with high Mg/Ca and low U/Ca[40–44], and we observed this same coupled variability in our coral subsamples (Extended Data Fig. 6). However this microstructural signal is typically small compared to the size of the boron isotope signals seen in our record. We had expected the smaller subsamples to exhibit more scatter, owing to the potential influence of coral centres of calcification (COCs), but this is not shown in these data. There is in fact slightly more variability between larger chunks, perhaps owing to the increased chance of sampling some COC material. For our records we used coral pieces of ~1 mg. We also mitigated microstructural influences by taking multiple solid subsamples from each coral when possible (shown as open symbols). Two subsamples (at 1.6 kyr ago and 20.4 kyr ago) were rejected from our total set of 55 as having anomalous $\delta^{11}B$ values (~1‰ lighter than the mean of 3–4 other subsamples from that coral), possibly owing to the impact of COC material.

Boron isotope analysis. Solid coral samples were crushed to a grain size <1 mm using an agate pestle and mortar. Samples were then subjected to oxidative cleaning to remove organic matter following established protocols[45–48], using warm 1% hydrogen peroxide, buffered in 0.1 M NH_4OH, followed by leaching in 0.0001 M HNO_3, and dissolution in 0.075 M distilled HNO_3. Boron was purified from the sample matrix using column chromatography with the boron-specific ion exchange resin Amberlite 743[49–51].

Boron isotope composition was analysed by MC-ICPMS by sample-standard bracketing[47,48,51]. Early analyses used NH_3 to improve boron washout in the spray chamber[52], which reduces background signals to ~3% of the preceding sample within ~3 min[51]. More recent analyses used dilute HF[53], which reduces background signals to ~0.5% within ~3 min. In contrast to some previous work, where samples were analysed in pure 0.3 M HF[53], we add a small volume of concentrated HF to our sample following column elution in 0.5 M HNO_3, giving a solution of 0.5 M HNO_3 + 0.3 M HF. Similarly, bracketing standards (NIST 951) and instrument blank acid were analysed in 0.5 M HNO_3 + 0.3 M HF to ensure consistent mass bias and blank corrections, though beyond improved washout we did not find any significant influence of HF concentration: boric acid standards run as dummy samples with HF concentrations from 0 to 0.5 M HF bracketed against 951 in 0.5 M HNO_3 + 0.3 M HF all yielded identical boron isotope ratios. Carbonate standards (JCP, NIST RM8301C) passed through columns and run with NH_3 or HF also yielded identical values.

All preparation and analytical work was carried out in boron-free clean laboratory conditions. Over the course of this work samples were analysed at the University of Bristol and Caltech on a Neptune MC-ICPMS, and the University of St Andrews on a Neptune Plus MC-ICPMS, though in all cases following nearly identical protocols. Each of these laboratories has taken part in published and ongoing inter-laboratory comparison studies[48] and there is no analytical offset between samples run in these laboratories.

Long-term analytical reproducibility on $\delta^{11}B$ measurements in these laboratories (assessed with carbonate standards given the same treatment as samples) is around 0.23 ‰ (2 s.d.) on samples of the size used here (~20 ng boron)—this is the error bar given on individual subsamples (open symbols in figures). We use a more conservative uncertainty for our mean coral $\delta^{11}B$ values (filled symbols in figures), to account for the $\delta^{11}B$ variability between subsamples from the same coral. This is based on the pooled s.d. of the replicate samples in our records and in Extended Data Fig. 6 (2 s.d. = 0.51 ‰). Uncertainty is reduced on mean values with multiple replicates (n), which is accounted for using the standard error (s.d./$\sqrt{n}$). Data from individual coral subsamples (open symbols in figures) are given in Extended Data Table 1 and average values from a given coral specimen (closed symbols and lines in figures) are given in Extended Data Table 2.

$\delta^{11}B$ and pH in deep-sea corals. The boron isotope pH proxy provides a sensitive measure of the ocean carbonate system, though in common with many proxies, is also influenced by modification during biomineralization[54–57]. In particular, coral $\delta^{11}B$ is influenced by internal pH elevation during biomineralization[58,59], which may buffer its sensitivity to external seawater pH changes in some settings. To examine this we have compiled modern *D. dianthus* $\delta^{11}B$ calibration data[44,56,60] from water depths >100 m, and have added two recent Southern Ocean samples (Extended Data Table 3). This indicates that the relationship between seawater pH and coral $\delta^{11}B$ is curved, with $\delta^{11}B$ becoming more sensitive to external seawater at lower pH (Extended Data Fig. 5). This suggests that corals find it harder to elevate internal pH as external conditions become more acidic, which is reasonable given the increase in energy demand required for additional proton expulsion[61]. It also means that corals are likely to be more sensitive to external pH conditions at lower pH sites, such as the Southern Ocean locations in this study. Hence, the pH changes that occurred during the last 40,000 years are evident as large, easily resolvable changes in $\delta^{11}B$ in our coral record.

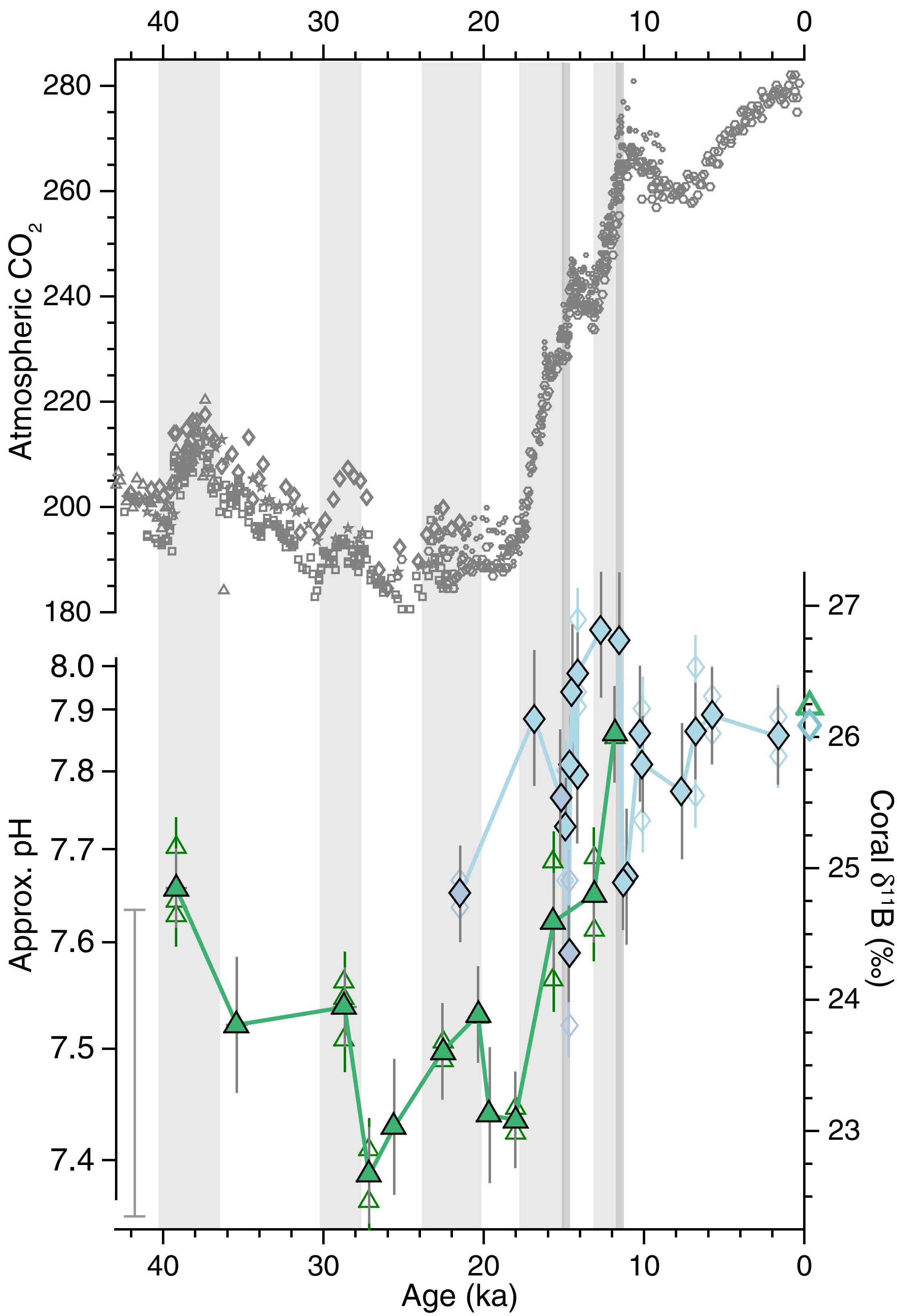
Reconstructed pH based on this calibration is shown in Extended Data Fig. 1. Given the paucity of modern deep-sea coral samples from low-pH waters, our record extends beyond the available calibration range. As a result, conversion to absolute pH values carries relatively large uncertainty, which is hard to assess. We thus prefer to focus on relative changes in the $\delta^{11}B$ records themselves, which provide a proxy of carbonate chemistry in their own right[54], analogous to the typical use of $\delta^{18}O$ records in paleoceanography.

Data availability

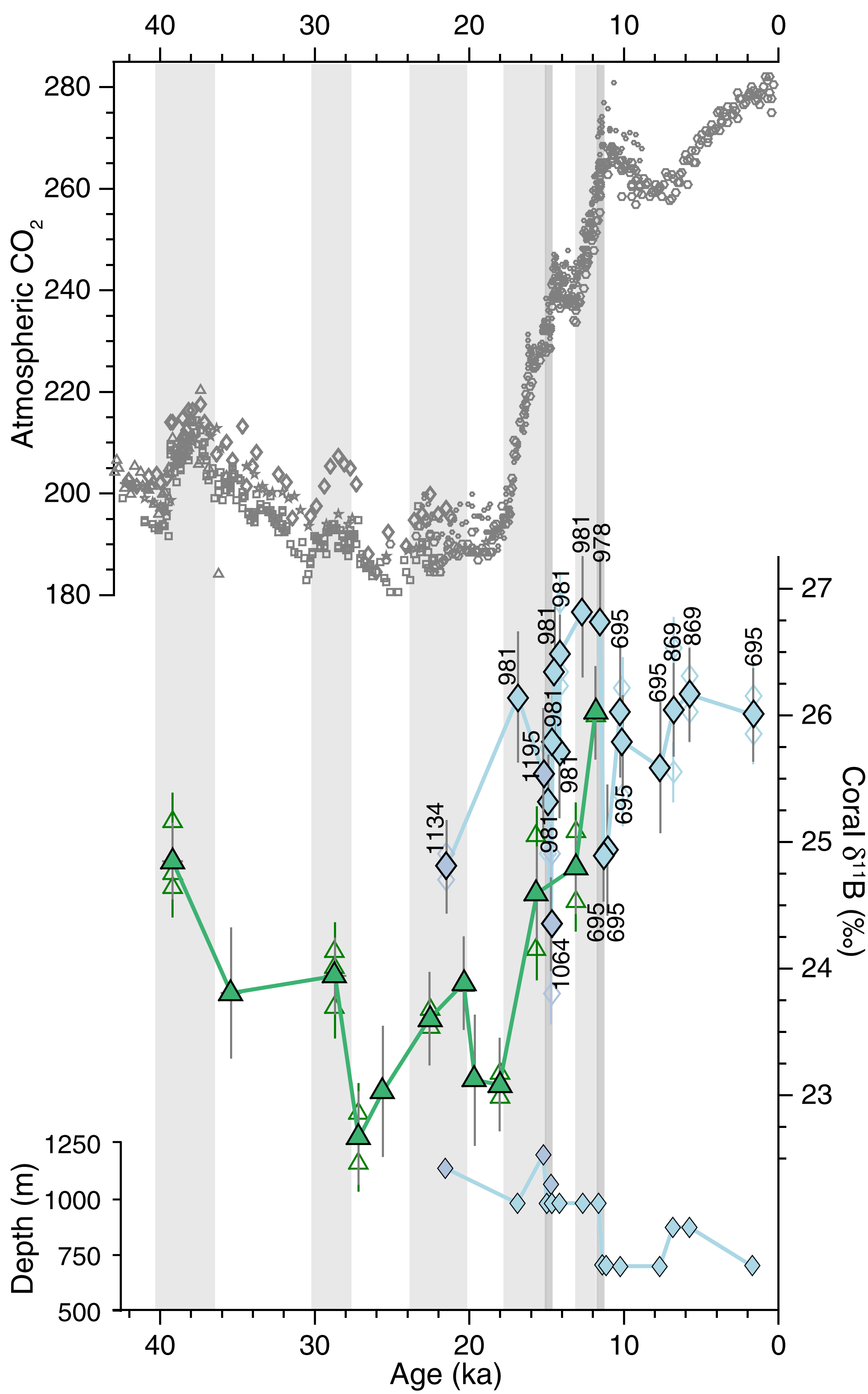
The data produced in this study are available in Extended Data Tables and will also be made available at the NOAA (https://www.ncdc.noaa.gov/paleo/study/25230) and Pangaea data repositories.

37. Burke, A. et al. Reconnaissance dating: A new radiocarbon method applied to assessing the temporal distribution of Southern Ocean deep-sea corals. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **57**, 1510–1520 (2010).
38. Margolin, A. R. et al. Temporal and spatial distributions of cold-water corals in the Drake Passage: Insights from the last 35,000 years. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **99**, 237–248 (2014).
39. Spooner, P. T., Chen, T., Robinson, L. F. & Coath, C. Rapid uranium-series age screening of carbonates by laser ablation mass spectrometry. *Quat. Geochronol.* **31**, 28–39 (2016).
40. Sinclair, D. J., Kinsley, L. P. & McCulloch, M. T. High resolution analysis of trace elements in corals by laser ablation ICP-MS. *Geochim. Cosmochim. Acta* **62**, 1889–1901 (1998).
41. Robinson, L. F. et al. Primary U distribution in scleractinian corals and its implications for U series dating. *Geochem. Geophys. Geosyst.* **7**, Q05022 (2006).
42. Gagnon, A. C., Adkins, J. F., Fernandez, D. P. & Robinson, L. F. Sr/Ca and Mg/Ca vital effects correlated with skeletal architecture in a scleractinian deep-sea coral and the role of Rayleigh fractionation. *Earth Planet. Sci. Lett.* **261**, 280–295 (2007).
43. Rollion-Bard, C., Chaussidon, M. & France-Lanord, C. Biological control of internal pH in scleractinian corals: Implications on paleo-pH and paleo-temperature reconstructions. *C. R. Geosci.* **343**, 397–405 (2011).
44. Stewart, J. A., Anagnostou, E. & Foster, G. L. An improved boron isotope pH proxy calibration for the deep-sea coral *Desmophyllum dianthus* through sub-sampling of fibrous aragonite. *Chem. Geol.* **447**, 148–160 (2016).
45. Boyle, E. A. Cadmium, zinc, copper, and barium in foraminifera tests. *Earth Planet. Sci. Lett.* **53**, 11–35 (1981).
46. Barker, S., Greaves, M. & Elderfield, H. A study of cleaning procedures used for foraminiferal Mg/Ca paleothermometry. *Geochem. Geophys. Geosyst.* **4**, 8407 (2003).
47. Rae, J. W. B., Foster, G. L., Schmidt, D. N. & Elliott, T. Boron isotopes and B/Ca in benthic foraminifera: Proxies for the deep ocean carbonate system. *Earth Planet. Sci. Lett.* **302**, 403–413 (2011).
48. Foster, G. L. et al. Interlaboratory comparison of boron isotope analyses of boric acid, seawater and marine $CaCO_3$ by MC-ICPMS and NTIMS. *Chem. Geol.* **358**, 1–14 (2013).
49. Kiss, E. Ion-exchange separation and spectrophotometric determination of boron in geological materials. *Anal. Chim. Acta* **211**, 243–256 (1988).
50. Lemarchand, D., Gaillardet, J., Göpel, C. & Manhès, G. An optimized procedure for boron separation and mass spectrometry analysis for river samples. *Chem. Geol.* **182**, 323–334 (2002).
51. Foster, G. L. Seawater pH, pCO_2 and $[CO_3^{=}]$ variations in the Caribbean Sea over the last 130 kyr: A boron isotope and B/Ca study of planktic foraminifera. *Earth Planet. Sci. Lett.* **271**, 254–266 (2008).
52. Al-Ammar, A. S., Gupta, R. K. & Barnes, R. M. Elimination of boron memory effect in inductively coupled plasma-mass spectrometry by ammonia gas injection into the spray chamber during analysis. *Spectrochim. Acta B At. Spectrosc.* **55**, 629–635 (2000).
53. Misra, S., Owen, R., Kerr, J. & Greaves, M. Determination of $\delta^{11}B$ by HR-ICP-MS from mass limited samples: application to natural carbonates and water samples. *Geochim. Cosmochim. Acta* **140**, 531–552 (2014).

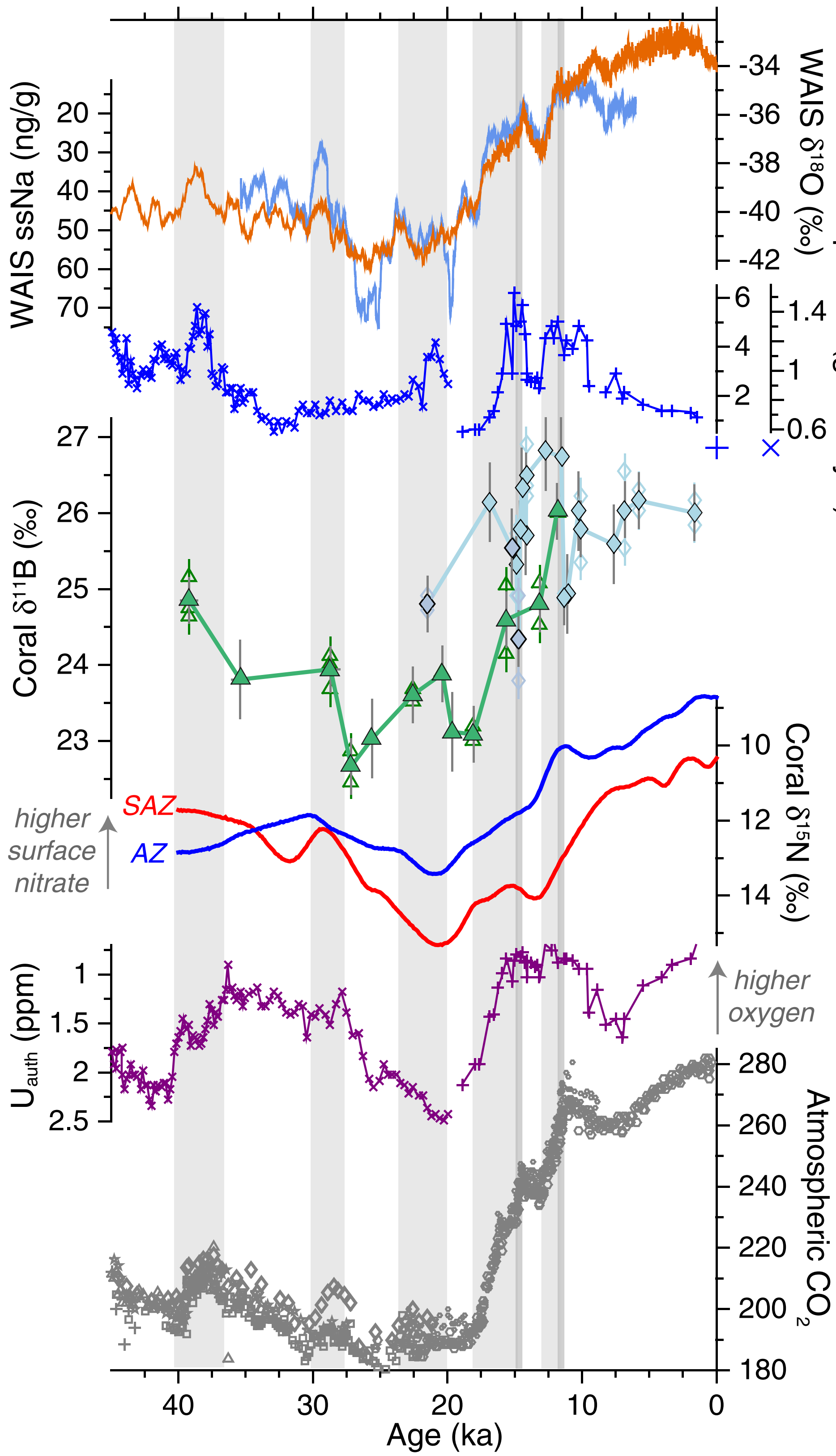
54. Rae, J. W. B. in *Boron Isotopes* 107–143 (Springer, 2018).
55. McCulloch, M. T. et al. in *Boron Isotopes* 145–162 (Springer, 2018).
56. Anagnostou, E., Huang, K. F., You, C. F., Sikes, E. L. & Sherrell, R. M. Evaluation of boron isotope ratio as a pH proxy in the deep sea coral *Desmophyllum dianthus*: evidence of physiological pH adjustment. *Earth Planet. Sci. Lett.* **349-350**, 251–260 (2012).
57. Trotter, J. et al. Quantifying the pH 'vital effect' in the temperate zooxanthellate coral *Cladocora caespitosa*: validation of the boron seawater pH proxy. *Earth Planet. Sci. Lett.* **303**, 163–173 (2011).
58. Venn, A. A. et al. Impact of seawater acidification on pH at the tissue-skeleton interface and calcification in reef corals. *Proc. Natl Acad. Sci. USA* **110**, 1634–1639 (2013).
59. Allison, N., Cohen, I., Finch, A. A., Erez, J. & Tudhope, A. W. Corals concentrate dissolved inorganic carbon to facilitate calcification. *Nat. Commun.* **5**, 5741 (2014).
60. McCulloch, M. et al. Resilience of cold-water scleractinian corals to ocean acidification: Boron isotopic systematics of pH and saturation state up-regulation. *Geochim. Cosmochim. Acta* **87**, 21–34 (2012).
61. Gagnon, A. C., Adkins, J. F., Erez, J. & Eiler, J. M. Sr/Ca sensitivity to aragonite saturation state in cultured subsamples from a single colony of coral: mechanism of biomineralization during ocean acidification. *Geochim. Cosmochim. Acta* **105**, 240–254 (2013).
62. Wang, X. T. et al. Deep-sea coral evidence for lower Southern Ocean surface nitrate concentrations during the last ice age. *Proc. Natl Acad. Sci. USA* **114**, 3352–3357 (2017).



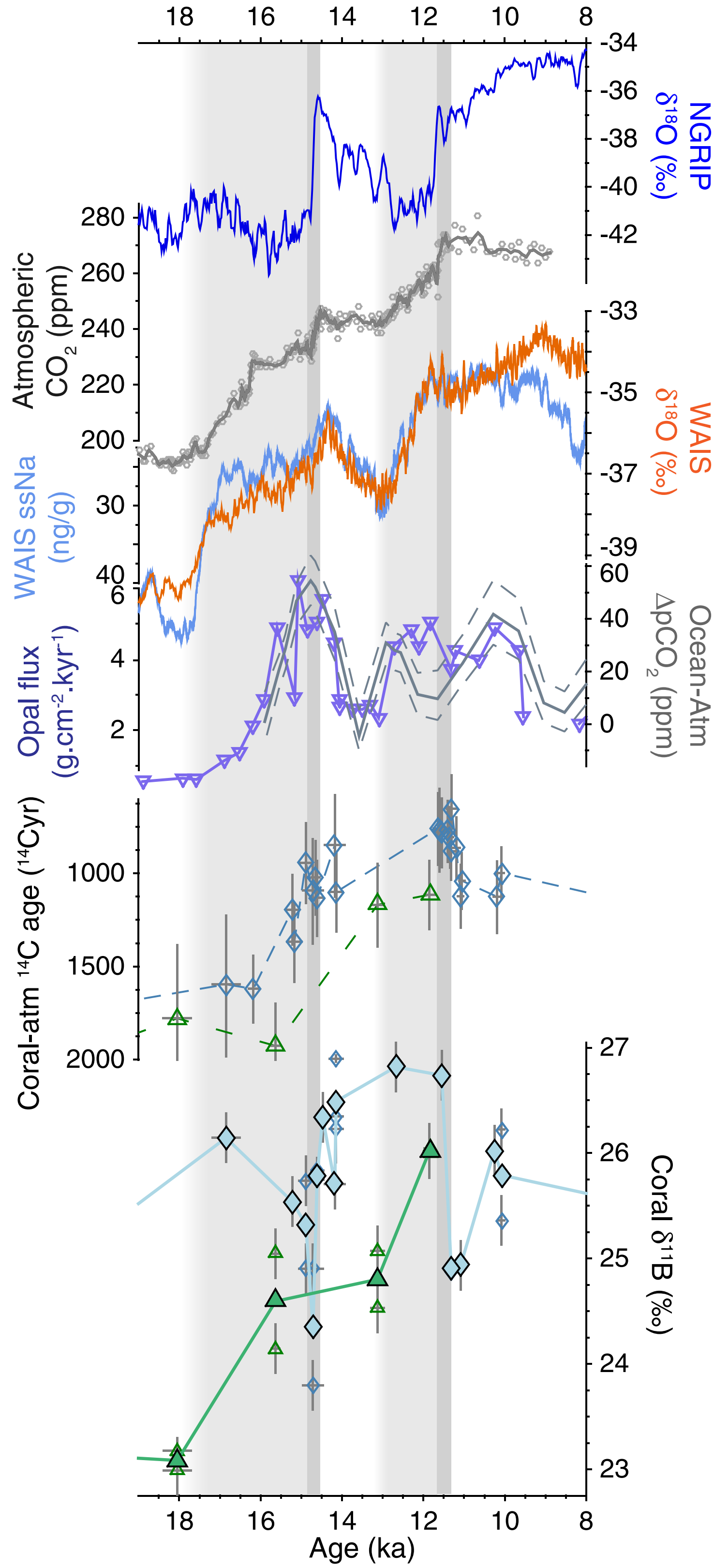
Extended Data Fig. 1 | Deep Southern Ocean CO_2 chemistry and atmospheric CO_2 over the last 40,000 years. Green triangles and blue diamonds show lower and upper cell deep-sea coral $\delta^{11}B$ data, respectively. Individual subsamples are shown as small open symbols and mean values as larger filled symbols. Error bars on individual subsamples are equivalent to 2 s.d. analytical reproducibility and error bars on mean coral values represent 2 s.e. uncertainty on the mean of replicate subsamples (see Methods). Approximate pH values are given based on coral $\delta^{11}B$ using the calibration in Extended Data Fig. 5, but uncertainty on this calibration is large (inset error bar), given the paucity of modern deep-sea coral data from low pH waters. Instead we focus on the $\delta^{11}B$ values themselves, which provide a proxy of carbonate chemistry in their own right[54]. Synchronized ice core CO_2 data[36] are shown in grey symbols: circles from Dome C, dots from WAIS, stars from Taylor Dome, triangles from TALDICE, pluses from EDML, diamonds from Byrd, and squares from Siple Dome. Grey bands highlight intervals of CO_2 rise.



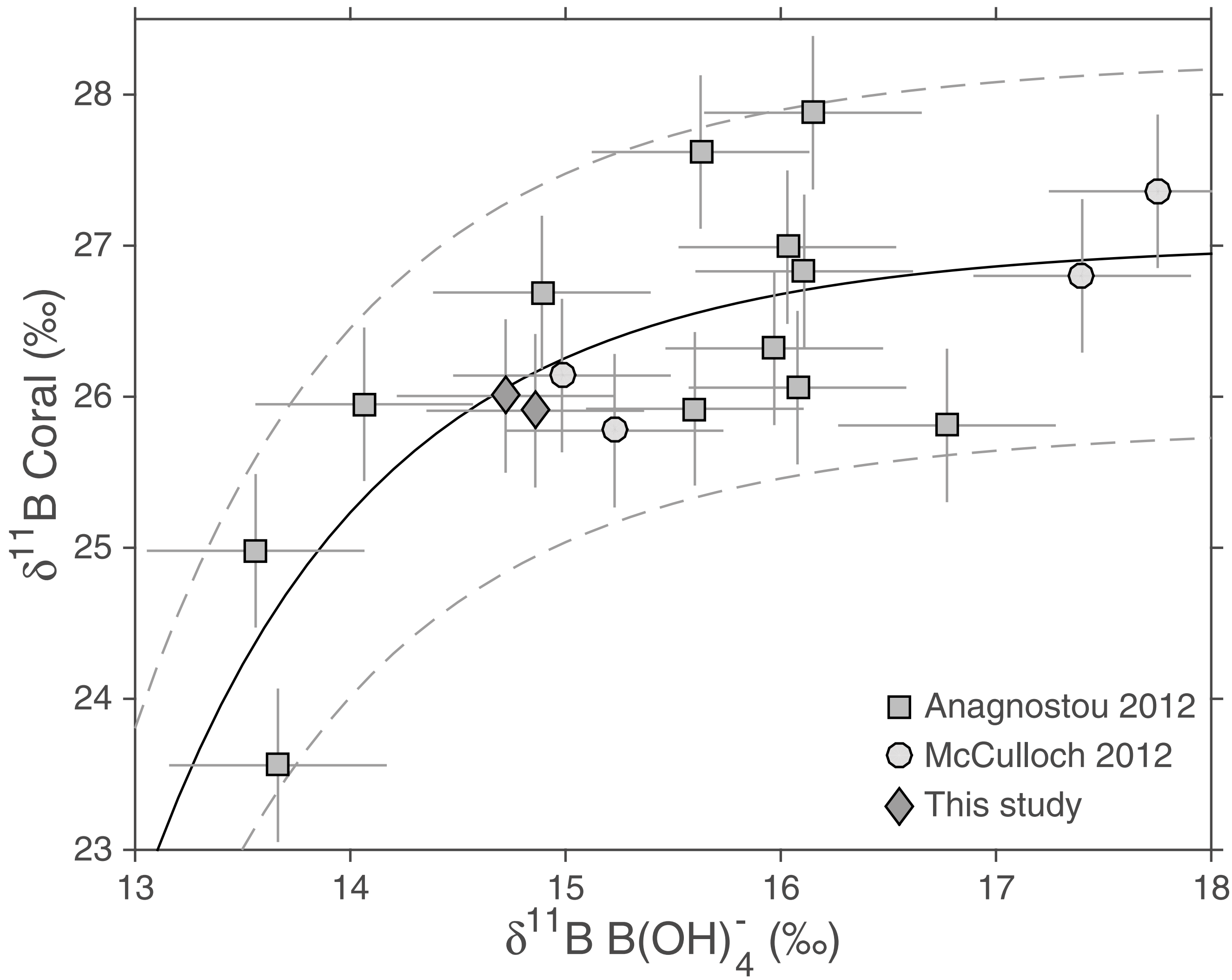
Extended Data Fig. 2 | Deep Southern Ocean CO_2 chemistry and atmospheric CO_2 over the last 40,000 years, highlighting the depths of upper cell corals. Symbols and data are as plotted in Extended Data Fig. 1, but with the addition of the lower panel and annotations showing the depth in metres of each upper cell coral sample. No systematic offset is seen between samples from different depths. The only signal that occurs simultaneous with a change in depth is the decrease at ~11.5 kyr ago, but the jump back up to higher $\delta^{11}B$ values following this event occurs without a change in depth, giving confidence that the excursion is not a depth-related signal. Furthermore the large excursion at ~14.7 kyr ago occurs without a significant change in depth. Note that all of the lower cell corals come from within 17 m water depth of each other.



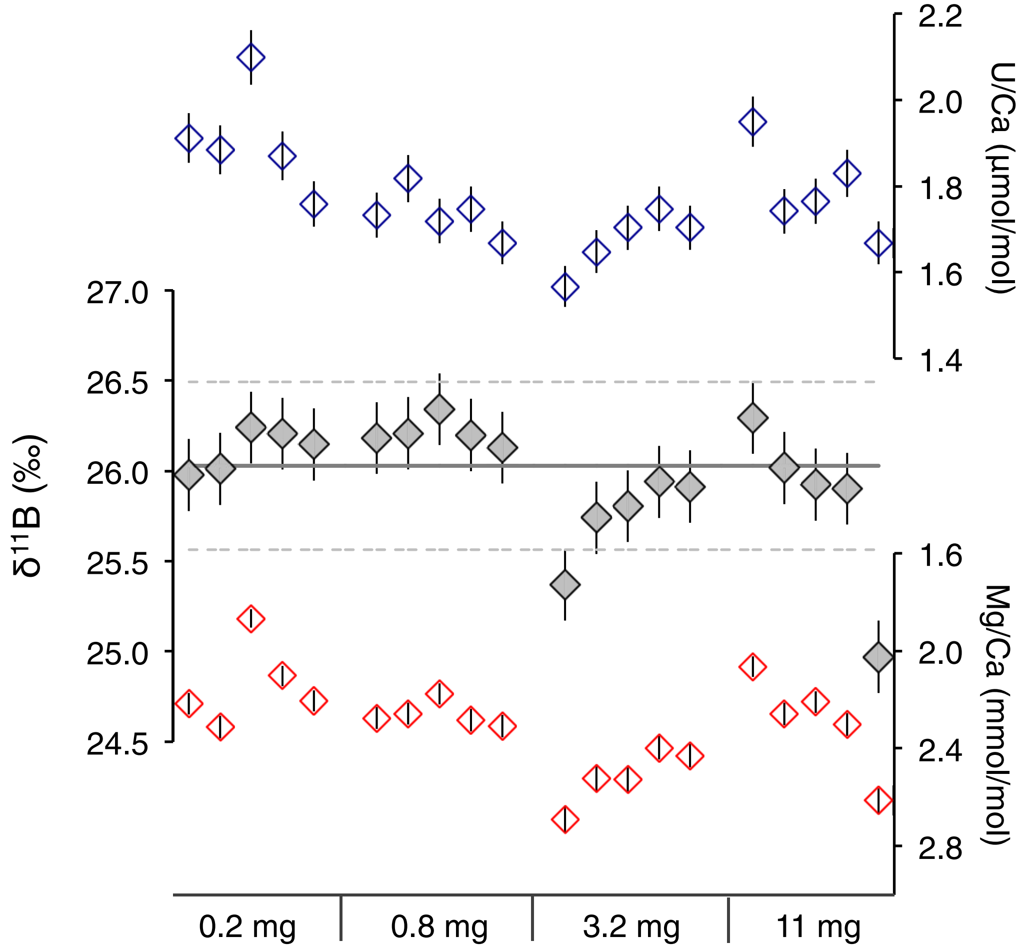
Extended Data Fig. 3 | Records of Southern Ocean biogeochemistry and CO_2 over the last 40,000 years. Data are plotted as in Fig. 2, but with opal flux[23], a proxy for upwelling, deep sea coral $\delta^{15}N$[62], a proxy for surface ocean nitrate consumption, and authigenic uranium concentrations[16], a proxy for bottom water redox. The opal flux and authigenic uranium records combine two sediment cores: TN057-13-4PC in the younger part of the record (pluses) and TN057-14PC in the older part of the record (crosses). The opal flux records from each core are shown on separate scales. The coral $\delta^{15}N$ data are grouped into samples from the Antarctic zone (AZ, blue) and Subantarctic zone (SAZ, red); smoothed fits to the data are shown, as provided in the original study[62]. Intervals of low CO_2 during the last ice age are associated with low upwelling, an efficient biological pump, low oxygen water rich in respired carbon, and low-pH carbon-rich water in the deep Southern Ocean.



Extended Data Fig. 4 | Deglacial records of Southern Ocean CO_2 chemistry and opal fluxes, and climate over Antarctica and Greenland. Data are plotted as in Figs. 2, 3, but with opal flux[23], a proxy for upwelling, surface ocean-atmosphere CO_2 difference, based on $\delta^{11}B$ in planktic foraminifera[26], and radiocarbon data[4,25] from corals within these sample groupings, shown as ^{14}C age offsets compared to the contemporaneous atmosphere. Intervals of rising CO_2 in the atmosphere are associated with input of waters rich in CO_2 and nutrients to the upper reaches of the Southern Ocean. Radiocarbon ages reflect the competing influences of upwelling of ^{14}C-depleted waters and improved ventilation over the deglaciation.



Extended Data Fig. 5 | Boron isotope calibration for modern *D. dianthus*. Data are from open ocean sites[44,56,60], with two additional recent (<1,650 years ago) samples from the Southern Ocean from this study. Water column δ^{11}B of borate ($B(OH)_4^-$) values are as previously published or are calculated from carbonate chemistry data from nearby GLODAPv2 sites for the new samples, as described[47,54]. Note that the sensitivity of δ^{11}B in carbonates to pH is based on the pH sensitivity of δ^{11}B of borate. pH itself is not easily shown on a plot like this, as the relationship between δ^{11}B of borate and pH is also somewhat influenced by water temperature, salinity, and depth[54]. A power law function was fitted to the data using Matlab's curve fitting toolbox (solid line: $\delta^{11}B_{Coral} = -1.82^{14} \times \delta^{11}B_{B(OH)4-}{}^{-12.22} + 27.03$; $R^2 = 0.57$). Dashed lines show the 95% confidence intervals and give a measure of calibration uncertainty as shown in the error bar in Extended Data Fig. 1, although data from a given site may be able to record relative changes in pH more sensitively, as seen in many paleo-proxies.



Extended Data Fig. 6 | Replicate subsamples from a *D. dianthus* septum. To test for the potential influence of microstructural variability in composition, a coral septum was divided into four areas, which were then split into chunks of approximately 0.2, 0.8, 3.2 and 11 mg. These were then individually crushed, cleaned, and analysed. This sample treatment was designed to preserve heterogeneity between subsamples, although note that the clustering of subsamples of a given size from a certain area of the coral may lead to that group recording a slightly different signal (as seen in the 3.2 mg group). The lines in the middle panel show the mean and 2 s.d., excluding one outlier in the 11 mg group. $\delta^{11}B$ is correlated with Mg/Ca and U/Ca, showing the influence of internal variability in coral composition.

Extended Data Table 1 | δ11B data for all fossil *D. dianthus* coral samples and subsamples (open symbols in Figures)

Cruise	Dredge	Sample	Latitude (°N)	Longitude (°E)	Depth (m)	Group	Location	Age (yrBP)	Age error (yr2SD)	$\delta^{11}B$ (‰)	$\delta^{11}B$ error (‰ 2SD)
NBP0805	DR23	Dc-A-7	-60.182	-57.828	819	LowerCell	SFZ	11850	62	26.05	0.23
NBP0805	DR23	Dc-A-7	-60.182	-57.828	819	LowerCell	SFZ	11850	62	25.99	0.23
NBP0805	DR23	Dc-A-5	-60.182	-57.828	819	LowerCell	SFZ	13127	77	24.53	0.23
NBP0805	DR23	Dc-A-5	-60.182	-57.828	819	LowerCell	SFZ	13127	77	25.07	0.23
NBP0805	DR23	Dc-A-6	-60.182	-57.828	819	LowerCell	SFZ	15629	49	24.15	0.23
NBP0805	DR23	Dc-A-6	-60.182	-57.828	819	LowerCell	SFZ	15629	49	25.04	0.23
NBP1103	DH40	Dc-3	-60.179	-57.837	806	LowerCell	SFZ	18047	168	22.99	0.23
NBP1103	DH40	Dc-3	-60.179	-57.837	806	LowerCell	SFZ	18047	168	23.18	0.23
NBP1103	DH43	Dc-6	-60.179	-57.003	823	LowerCell	SFZ	19631	189	23.12	0.23
NBP1103	DH43	Dc-1	-60.179	-57.003	823	LowerCell	SFZ	20367	58	23.87	0.23
NBP1103	DH43	Dc-1	-60.179	-57.003	823	LowerCell	SFZ	20367	58	23.90	0.23
NBP1103	DH40	Dc-5	-60.179	-57.837	806	LowerCell	SFZ	22578	81	23.68	0.23
NBP1103	DH40	Dc-5	-60.179	-57.837	806	LowerCell	SFZ	22578	81	23.53	0.23
NBP0805	DR23	Dc-A-1	-60.182	-57.828	819	LowerCell	SFZ	25587	160	23.03	0.23
NBP1103	DH43	Dc-3	-60.179	-57.003	823	LowerCell	SFZ	27159	122	22.46	0.23
NBP1103	DH43	Dc-3	-60.179	-57.003	823	LowerCell	SFZ	27159	122	22.86	0.23
NBP1103	DH43	Dc-12	-60.179	-57.003	823	LowerCell	SFZ	28679	329	24.01	0.23
NBP1103	DH43	Dc-12	-60.179	-57.003	823	LowerCell	SFZ	28679	329	24.13	0.23
NBP1103	DH43	Dc-12	-60.179	-57.003	823	LowerCell	SFZ	28679	329	23.69	0.23
NBP0805	DR23	Dc-A-2	-60.182	-57.828	819	LowerCell	SFZ	35406	306	23.81	0.23
NBP0805	DR23	Dc-A-3	-60.182	-57.828	819	LowerCell	SFZ	39193	301	25.15	0.23
NBP0805	DR23	Dc-A-3	-60.182	-57.828	819	LowerCell	SFZ	39193	301	24.76	0.23
NBP0805	DR23	Dc-A-3	-60.182	-57.828	819	LowerCell	SFZ	39193	301	24.64	0.23
NBP0805	DR35	Dc-A-1	-59.721	-68.883	695	UpperCell	Sars	1649	13	26.16	0.23
NBP0805	DR35	Dc-A-1	-59.721	-68.883	695	UpperCell	Sars	1649	13	25.85	0.23
NBP0805	DR34	Dc-A-2	-59.732	-68.746	869	UpperCell	Sars	5759	30	26.30	0.23
NBP0805	DR34	Dc-A-2	-59.732	-68.746	869	UpperCell	Sars	5759	30	26.02	0.23
NBP0805	DR34	Dc-A-1	-59.732	-68.746	869	UpperCell	Sars	6798	34	26.54	0.23
NBP0805	DR34	Dc-A-1	-59.732	-68.746	869	UpperCell	Sars	6798	34	25.55	0.23
NBP0805	DR35	Dc-C-2	-59.721	-68.883	695	UpperCell	Sars	7639	28	25.59	0.23
NBP0805	DR35	Dc-D-4	-59.721	-68.883	695	UpperCell	Sars	10084	34	25.36	0.23
NBP0805	DR35	Dc-D-4	-59.721	-68.883	695	UpperCell	Sars	10084	34	26.22	0.23
NBP0805	DR35	Dc-D5	-59.721	-68.883	695	UpperCell	Sars	10255	81	26.03	0.23
NBP0805	DR35	Dc-D-3	-59.721	-68.883	695	UpperCell	Sars	11084	34	24.94	0.23
NBP0805	DR35	Dc-B-1a	-59.721	-68.883	695	UpperCell	Sars	11317	36	24.87	0.23
NBP0805	DR35	Dc-B-1a	-59.721	-68.883	695	UpperCell	Sars	11317	36	24.93	0.23
NBP0805	DR38	Dc-A-1	-59.741	-68.900	978	UpperCell	Sars	11549	64	26.74	0.23
NBP1103	DH117	Dn-1	-59.764	-68.936	981	UpperCell	Sars	12680	102	26.82	0.23
NBP1103	DH117	Dn-7	-59.764	-68.936	981	UpperCell	Sars	14143	80	26.90	0.23
NBP1103	DH117	Dn-7	-59.764	-68.936	981	UpperCell	Sars	14143	80	26.35	0.23
NBP1103	DH117	Dn-7	-59.764	-68.936	981	UpperCell	Sars	14143	80	26.23	0.23
NBP1103	DH117	Dc-20	-59.764	-68.936	981	UpperCell	Sars	14177	122	25.71	0.23
NBP1103	DH117	Dc-1	-59.764	-68.936	981	UpperCell	Sars	14475	111	26.34	0.23
NBP1103	DH117	Dc-29	-59.764	-68.936	981	UpperCell	Sars	14609	74	25.83	0.23
NBP1103	DH117	Dc-29	-59.764	-68.936	981	UpperCell	Sars	14609	74	25.75	0.23
NBP1103	DH74	Dc-3	-60.606	-66.004	1064	UpperCell	Interim	14714	122	23.80	0.23
NBP1103	DH74	Dc-3	-60.606	-66.004	1064	UpperCell	Interim	14714	122	24.90	0.23
NBP1103	DH117	Dc-36	-59.764	-68.936	981	UpperCell	Sars	14889	80	24.90	0.23
NBP1103	DH117	Dc-36	-59.764	-68.936	981	UpperCell	Sars	14889	80	25.74	0.23
NBP1103	DH75	Dc(f)-37	-60.613	-66.002	1196	UpperCell	Interim	15221	72	25.54	0.23
NBP1103	DH117	Dc-09	-59.764	-68.936	981	UpperCell	Sars	16851	168	26.14	0.23
NBP0805	DR27	Dc-A-1	-60.546	-65.953	1134	UpperCell	Interim	21466	109	24.90	0.23
NBP0805	DR27	Dc-A-1	-60.546	-65.953	1134	UpperCell	Interim	21466	109	24.71	0.23

This table is also available as a spreadsheet in the Supplementary Information and on Pangaea and NCDC. Location groupings are: SFZ, Shackleton fracture zone; Sars, Sars Seamount; Interim, Interim Seamount (see map in Fig. 1).

Extended Data Table 2 | Averaged $\delta^{11}B$ data from each *D. dianthus* coral specimen (filled symbols in figures)

Cruise	Dredge	Sample	Latitude (°N)	Longitude (°E)	Depth (m)	Group	Location	Age (yrBP)	Age error (yr2SD)	$\delta^{11}B$ (‰)	n	$\delta^{11}B$ error (‰ 2SD)
NBP0805	DR23	Dc-A-7	-60.182	-57.828	819	LowerCell	SFZ	11850	62	26.02	2	0.36
NBP0805	DR23	Dc-A-5	-60.182	-57.828	819	LowerCell	SFZ	13127	77	24.80	2	0.36
NBP0805	DR23	Dc-A-6	-60.182	-57.828	819	LowerCell	SFZ	15629	49	24.59	2	0.36
NBP1103	DH40	Dc-3	-60.179	-57.837	806	LowerCell	SFZ	18047	168	23.08	2	0.36
NBP1103	DH43	Dc-6	-60.179	-57.003	823	LowerCell	SFZ	19631	189	23.12	1	0.51
NBP1103	DH43	Dc-1	-60.179	-57.003	823	LowerCell	SFZ	20367	58	23.89	2	0.36
NBP1103	DH40	Dc-5	-60.179	-57.837	806	LowerCell	SFZ	22578	81	23.61	2	0.36
NBP0805	DR23	Dc-A-1	-60.182	-57.828	819	LowerCell	SFZ	25587	160	23.03	1	0.51
NBP1103	DH43	Dc-3	-60.179	-57.003	823	LowerCell	SFZ	27159	122	22.66	2	0.36
NBP1103	DH43	Dc-12	-60.179	-57.003	823	LowerCell	SFZ	28679	329	23.94	3	0.29
NBP0805	DR23	Dc-A-2	-60.182	-57.828	819	LowerCell	SFZ	35406	306	23.81	1	0.51
NBP0805	DR23	Dc-A-3	-60.182	-57.828	819	LowerCell	SFZ	39193	301	24.85	3	0.29
NBP0805	DR35	Dc-A-1	-59.721	-68.883	695	UpperCell	Sars	1649	13	26.00	2	0.36
NBP0805	DR34	Dc-A-2	-59.732	-68.746	869	UpperCell	Sars	5759	30	26.16	2	0.36
NBP0805	DR34	Dc-A-1	-59.732	-68.746	869	UpperCell	Sars	6798	34	26.04	2	0.36
NBP0805	DR35	Dc-C-2	-59.721	-68.883	695	UpperCell	Sars	7639	28	25.59	1	0.51
NBP0805	DR35	Dc-D-4	-59.721	-68.883	695	UpperCell	Sars	10084	34	25.79	2	0.36
NBP0805	DR35	Dc-D5	-59.721	-68.883	695	UpperCell	Sars	10255	81	26.03	1	0.51
NBP0805	DR35	Dc-D-3	-59.721	-68.883	695	UpperCell	Sars	11084	34	24.94	1	0.51
NBP0805	DR35	Dc-B-1a	-59.721	-68.883	695	UpperCell	Sars	11317	36	24.90	2	0.36
NBP0805	DR38	Dc-A-1	-59.741	-68.900	978	UpperCell	Sars	11549	64	26.74	1	0.51
NBP1103	DH117	Dn-1	-59.764	-68.936	981	UpperCell	Sars	12680	102	26.82	1	0.51
NBP1103	DH117	Dn-7	-59.764	-68.936	981	UpperCell	Sars	14143	80	26.49	3	0.29
NBP1103	DH117	Dc-20	-59.764	-68.936	981	UpperCell	Sars	14177	122	25.71	1	0.51
NBP1103	DH117	Dc-1	-59.764	-68.936	981	UpperCell	Sars	14475	111	26.34	1	0.51
NBP1103	DH117	Dc-29	-59.764	-68.936	981	UpperCell	Sars	14609	74	25.79	2	0.36
NBP1103	DH74	Dc-3	-60.606	-66.004	1064	UpperCell	Interim	14714	122	24.35	2	0.36
NBP1103	DH117	Dc-36	-59.764	-68.936	981	UpperCell	Sars	14889	80	25.32	2	0.36
NBP1103	DH75	Dc(f)-37	-60.613	-66.002	1196	UpperCell	Interim	15221	72	25.54	1	0.51
NBP1103	DH117	Dc-09	-59.764	-68.936	981	UpperCell	Sars	16851	168	26.14	1	0.51
NBP0805	DR27	Dc-A-1	-60.546	-65.953	1134	UpperCell	Interim	21466	109	24.81	2	0.36

This table is also available as a spreadsheet in the Supplementary Information and on Pangaea. Location groupings are: SFZ, Shackleton fracture zone; Sars, Sars Seamount; Interim, Interim Seamount (see map in Fig. 1). Uncertainty on mean coral $\delta^{11}B$ values is based on the pooled s.d. of all the replicates in this study (including those shown in Extended Data Fig. 6), divided by square root of *n*, to give a measure of the s.e.

Extended Data Table 3 | *D. dianthus* $\delta^{11}B$ calibration data (as shown in Extended Data Fig. 5)

Study	Sample	Depth (m)	Latitude (°N)	Longitude (°E)	T (°C)	S (psu)	ALK (µmol/kg)	DIC (µmol/kg)	pH total scale	$\Omega_{aragonite}$	$\delta^{11}B_{borate}$ (‰)	$\delta^{11}B_{coral}$ (‰)	$\delta^{11}B_{coral}$ error (‰ 2SD)	$\delta^{11}B_{coral}$ subsamples (‰)
McCulloch et al., 2012	Tasman Seamount DD_MS	932	40.76	29.17	14.5	38.8	2610	2469	7.77	1.46	15.23	25.78	0.31	25.68, 25.87
McCulloch et al., 2012	Marmara Sea Hill_B1	1050	-44.33	147.28	4.59	34.4	2315	2217	7.87	1.02	14.98	26.14	0.31	
McCulloch et al., 2012	MedCor-25-D	576	35.52	14.18	13.78	38.75	2613	2314	8.1	2.88	17.75	27.36	0.31	
McCulloch et al., 2012	MedCor-74-D	837	36.75	13.99	13.96	38.77	2624	2347	8.05	2.59	17.40	26.8	0.31	
Anagnostou et al., 2012	19249	274	34.00	-119.50	7.8	34.11	2271	2226	7.69	0.83	14.06	25.95	0.25	25.89, 26.08
Anagnostou et al., 2012	47409	673	−54.5	-39.40	1.8	34.69	2336	2238	7.91	1.08	14.89	26.69	0.25	26.63, 26.82
Anagnostou et al., 2012	47413	421	−50.6	167.60	5.9	34.26	2284	2124	8.05	1.63	16.11	26.83	0.25	26.81, 26.89
Anagnostou et al., 2012	48473	1107	47.70	-8.10	8.5	35.55	2348	2182	7.97	1.5	16.08	26.06	0.25	
Anagnostou et al., 2012	48739	825	47.60	-7.30	9.6	35.53	2344	2172	7.98	1.64	16.15	27.88	0.25	
Anagnostou et al., 2012	48740	1470	48.70	-10.90	5.1	35.12	2324	2169	7.99	1.28	16.03	26.99	0.25	
Anagnostou et al., 2012	62309	522	40.40	-67.70	6.1	35.05	2317	2179	7.97	1.43	15.60	25.92	0.25	25.82, 26.02
Anagnostou et al., 2012	78630	312	46.80	-130.80	5.8	33.96	2280	2265	7.61	0.65	13.66	23.56	0.25	
Anagnostou et al., 2012	80358	358	48.00	-7.90	11	35.54	2336	2129	8.06	2.11	16.77	25.81	0.25	
Anagnostou et al., 2012	83583	464	32.90	-127.80	5.8	34.1	2290	2284	7.57	0.59	13.56	24.98	0.25	24.95, 25.04
Anagnostou et al., 2012	94069	710	−30.5	-178.70	7.4	34.49	2284	2146	7.95	1.38	15.63	27.62	0.25	27.48, 27.89
Anagnostou et al., 2012	Z9725	276	−45.2	171.60	8.1	34.51	2288	2130	8.01	1.68	15.97	26.32	0.25	
Rae - this study	NBP0805_DR40_DCA1	1323	-59.73	-68.93	2.13	34.67	2347	2259	7.87	0.91	14.86	25.91	0.23	25.71, 25.95, 26.06
Rae - this study	NBP0805_DR35_DCA1	695	-59.72	-68.88	2.45	34.46	2320	2237	7.88	1.00	14.72	26.00	0.23	25.85, 26.16

This table is also available as a spreadsheet in the Supplementary Information and on Pangaea. Latitude, longitude, and depth values are given as averages where samples were collected by dredge. DIC in the McCulloch et al. (2011)[55] sample set has been calculated from reported ALK and pH. Carbonate system parameters in the Anagnostou et al. (2011)[56] sample set are as re-calculated in Stewart et al. (2016)[44], using GLODAP ALK and DIC. Carbonate system parameters for new data reported here are calculated from GLODAPv2 ALK and DIC. Coral subsamples represent different solid pieces/powder subsampled from the same specimen. Note that Anagnostou et al. (2011)[56] give a weighted average of these values, which is reported in the $\delta^{11}B_{coral}$ column.

LETTER

https://doi.org/10.1038/s41586-018-0613-1

Social regulation of a rudimentary organ generates complex worker-caste systems in ants

Rajendhran Rajakumar[1,2], Sophie Koch[1], Mélanie Couture[1], Marie-Julie Favé[1,3], Angelica Lillico-Ouachour[1], Travis Chen[1], Giovanna De Blasis[1], Arjuna Rajakumar[1], Dominic Ouellette[1] & Ehab Abouheif[1]*

The origin of complex worker-caste systems in ants perplexed Darwin[1] and has remained an enduring problem for evolutionary and developmental biology[2–6]. Ants originated approximately 150 million years ago, and produce colonies with winged queen and male castes as well as a wingless worker caste[7]. In the hyperdiverse genus *Pheidole*, the wingless worker caste has evolved into two morphologically distinct subcastes—small-headed minor workers and large-headed soldiers[8]. The wings of queens and males develop from populations of cells in larvae that are called wing imaginal discs[7]. Although minor workers and soldiers are wingless, vestiges or rudiments of wing imaginal discs appear transiently during soldier development[7,9–11]. Such rudimentary traits are phylogenetically widespread and are primarily used as evidence of common descent, yet their functional importance remains equivocal[1,12–14]. Here we show that the growth of rudimentary wing discs is necessary for regulating allometry—disproportionate scaling—between head and body size to generate large-headed soldiers in the genus *Pheidole*. We also show that *Pheidole* colonies have evolved the capacity to socially regulate the growth of rudimentary wing discs to control worker subcaste determination, which allows these colonies to maintain the ratio of minor workers to soldiers. Finally, we provide comparative and experimental evidence that suggests that rudimentary wing discs have facilitated the parallel evolution of complex worker-caste systems across the ants. More generally, rudimentary organs may unexpectedly acquire novel regulatory functions during development to facilitate adaptive evolution.

The evolution of large-headed soldiers is thought to have promoted the adaptive radiation of *Pheidole*, which has produced approximately 1,000 species worldwide[8,15]. *Pheidole* soldiers are larger than minor workers and their heads are disproportionally larger than their bodies, forming an allometric line that is distinct from that of minor workers (Fig. 1a–c). A developmental switch—largely controlled by nutrition[6] and mediated by juvenile hormone[16]—determines whether larvae develop into minor workers or soldiers (Fig. 2a). Minor worker larvae lack wing rudiments (Fig. 2c), whereas soldier larvae have one pair of rudimentary forewing discs[9] (Fig. 2d). In *Pheidole obtusospinosa*, which has evolved a disproportionately larger 'supersoldier' subcaste (Fig. 1d–f), supersoldier larvae have two pairs of large rudimentary wing discs[10] (Fig. 1g–i). In addition, using Pagel's[17] phylogenetic correlation method, we found that the size of rudimentary wing discs varies discretely between worker subcastes within ant species that have independently evolved a soldier subcaste (Extended Data Fig. 1 and Extended Data Table 1). Finally, rudimentary forewing discs in *Pheidole* soldier larvae are coordinated in their growth, gene expression and apoptosis—just after larvae become determined as soldiers (Fig. 2a, red arrowhead), rudimentary forewing discs appear[9], grow rapidly[9] (Extended Data Fig. 2), activate expression of genes in the wing network[7,10] and are finally eliminated by apoptosis[11]. These findings raise the possibility that rudimentary wing discs have a functional role in the development of soldiers.

We tested this possibility in *Pheidole hyatti* by targeting *vestigial* (*vg*), which in *Drosophila* is a selector gene that coordinates growth and patterning of wing imaginal discs and is necessary and sufficient for wing development[18–20]. Spatial expression of *vg* is similar in *P. hyatti* and *Drosophila*—in embryos *vg* is expressed in wing primordia and the ventral nerve cord, but in larvae *vg* expression could be detected only in the wing discs of winged castes and in the rudimentary forewing discs of soldiers (Fig. 2b–d and Extended Data Fig. 3a–o). We therefore used RNA-mediated interference (RNAi) to knockdown *vg* expression in soldier-destined larvae (Fig. 2a, red arrowhead). Compared to controls, *vg* RNAi significantly reduces the size of rudimentary forewing discs (Fig. 2e–g) and, as in *Drosophila* wing discs, perturbing *vg* expression induces apoptosis (Extended Data Fig. 4). We discovered that reducing the size of rudimentary forewing discs affects the adult phenotype: it significantly reduces the head/body ratio and changes the head-to-body slope such that head size is reduced more than body size, relative to controls (Fig. 2h, i and Extended Data Fig. 5a–f). These ants vary widely in head and body size, from being as small as minor workers to being as large as soldiers—some ants are intermediate in size, which are variants that do not exist in nature (Fig. 2i and Extended Data Fig. 5g–l). Finally, *vg* has also been shown to be expressed outside of wing discs in *Drosophila* larvae[21]. To confirm that *vg* RNAi changes head-to-body scaling by specifically reducing the size of rudimentary forewing discs, we electro-surgically ablated the left rudimentary forewing disc in soldier-destined larvae; as a control, we ablated the third leg disc on the left side (Fig. 2a, red arrowhead and Extended Data Fig. 5m–q). These ablations produced results similar to *vg* RNAi (Fig. 2j, k and Extended Data Fig. 5r–t). Together, these experiments show that rudimentary forewing discs are necessary to regulate the size and disproportionate head-to-body scaling that generates large-headed soldiers.

Minor worker larvae lack rudimentary wing discs and do not express *vg* in any other imaginal discs (Fig. 2c and Extended Data Fig. 3e, h, k, n). However, we found low levels of expression of *vg* outside of imaginal discs in minor worker larvae (Extended Data Fig. 3p). To test whether this expression influences head-to-body scaling, we treated bipotential larvae (Fig. 2a, orange arrowhead)—of which 90–95% develop into minor workers, and 5–10% into soldiers[16]—with *vg* RNAi. The bipotential larvae treated with *vg* RNAi that became minor workers did not show significant differences in size, head/body ratio, slope or intercept compared to controls (Extended Data Fig. 6a, b, e, g-i), whereas the few that developed into soldier-destined larvae became intermediates (Extended Data Fig. 6c, d, f). Therefore, expression of *vg* outside of imaginal discs does not influence size or head-to-body scaling in the minor-worker subcaste. We then investigated whether wing discs in winged castes regulate head-to-body scaling. Relative to controls, we found that in larvae destined to become males (Fig. 2a, green arrowhead) *vg* RNAi causes defects in patterning and growth of adult fore- and hindwings (Extended Data Fig. 6j–l), but has no significant effect on size, head/body ratio, slope or intercept (Extended Data Fig. 6m–p). Therefore, wing discs in male larvae produce wings

[1]Department of Biology, McGill University, Montreal, Quebec, Canada. [2]Present address: Department of Genetics, Harvard Medical School, Boston, MA, USA. [3]Present address: Ontario Institute for Cancer Research, Toronto, Ontario, Canada. *e-mail: ehab.abouheif@mcgill.ca

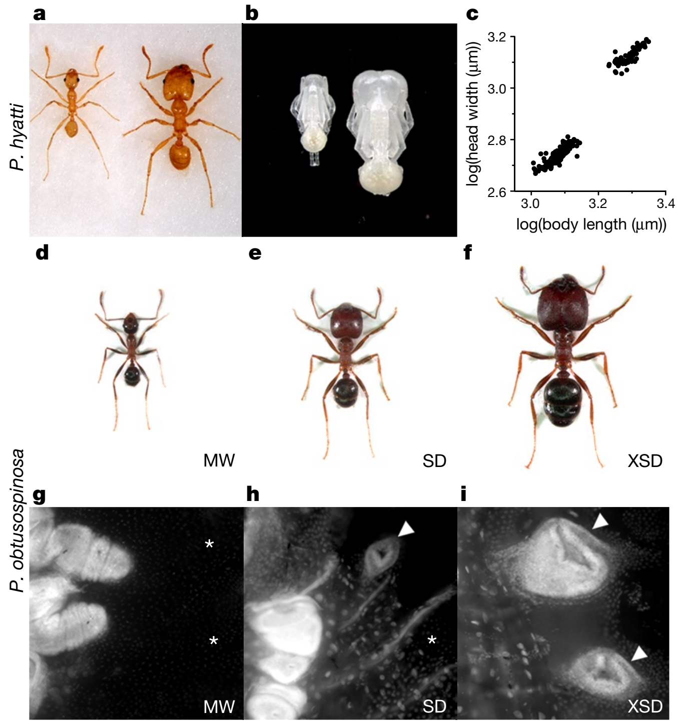


Fig. 1 | Head-to-body allometry and rudimentary wing discs of *Pheidole* worker subcastes. a, *P. hyatti* minor worker adult (left) and soldier adult (right). **b**, *P. hyatti* minor worker pupa (left) and soldier pupa (right). **c**, log–log plot of head-to-body allometry of wild-type minor workers ($n = 155$) and soldiers ($n = 80$) (x and y axes in µm). **d**–**i**, Comparing *P. obtusospinosa* worker subcastes and rudimentary wing discs (comparisons to scale): minor worker (MW, **d**, **g**), soldier (SD, **e**, **h**) and supersoldier (XSD, **f**, **i**). Arrowheads indicate rudimentary wing discs; asterisk indicates absence of a rudimentary wing disc.

but do not regulate size or head-to-body scaling. These results show that, within the wingless female worker caste, rudimentary forewing discs regulate the size and disproportionate head-to-body scaling that is specific to the soldier subcaste.

Pheidole colonies dynamically maintain a ratio of soldiers (5–10%) to minor workers (90–95%): soldiers use their large heads for defence and food processing, and minor workers forage and nurse young[8,22–25]. This ratio is maintained through the balance of two types of social regulation: (1) the activation of soldier development by minor workers through nutrition[6], which is mediated by juvenile hormone[16]; and (2) the suppression of soldier development through an inhibitory pheromone when the number of soldiers increases to above 5–10%[23,24]. This pheromone is contact-based and is thought to be composed of cuticular hydrocarbons[23,24,26]. Because juvenile hormone induces rudimentary forewing disc growth in *P. hyatti*[10], we investigated whether their growth is also influenced by the inhibitory pheromone. We therefore raised soldier-destined larvae (Fig. 2a, red arrowhead) with either 100% soldiers (high inhibition) or 100% minor workers (no inhibition, as a control). We discovered that raising soldier-destined larvae with 100% soldiers significantly reduces the size of rudimentary forewing discs, relative to controls (Fig. 3a–c). This produced ants with a significantly reduced head/body ratio and changes the head-to-body slope such that head size is reduced more than body size, relative to controls (Fig. 3d, e and Extended Data Fig. 7a–f). These ants vary widely in size, ranging from minor workers to soldiers, and include intermediates (Fig. 3e and Extended Data Fig. 7g–k). Finally, to rule out the possibility that these results are caused by any potential differences in the rearing environments of minor workers and soldiers, we treated soldier-destined larvae with cuticular hydrocarbons extracted from soldiers and compared them to solvent-only controls under identical rearing conditions (100% minor workers). This produced

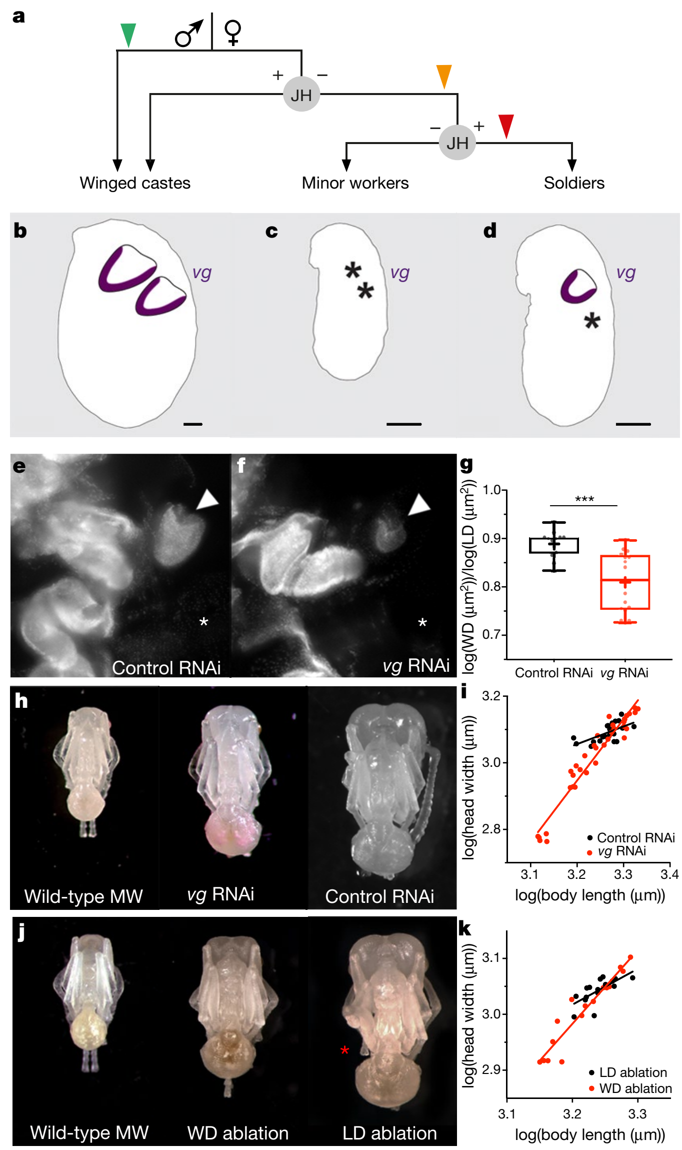


Fig. 2 | Rudimentary forewing discs regulate size and disproportionate head-to-body scaling of soldiers. a, Caste determination in *Pheidole* at three developmental switch points produces: winged males, winged queens, wingless minor workers and wingless soldiers. Points of experimental manipulation are indicated by coloured arrowheads: red, soldier-destined larvae; orange, bipotential larvae; green, male-destined larvae. JH, juvenile hormone. **b**–**d**, *vg* expression (purple) in larval wing discs of males or queens (**b**), minor workers (**c**) and soldiers (**d**). Black asterisk, absence of rudimentary wing disc. Scale bars, relative scale. **e**–**f**, Rudimentary forewing discs after *yfp* RNAi (control RNAi) (**e**) or *vg* RNAi (**f**). Rudimentary wing disc presence (white arrowheads) or absence (white asterisks). **g**, Comparing ratio of log(rudimentary forewing disc area (µm²)) to log(leg disc area (µm²)) (log(WD (µm²))/log(LD (µm²))) between control RNAi ($n = 13$) and *vg* RNAi ($n = 16$). The box plot shows mean (+), interquartile range (bars) and minimum to maximum values (whiskers); all points represent individual ants. Two-tailed Mann–Whitney U-test, $U = 24$, $***P = 0.0002$. **h**, Wild-type minor worker and representative individuals treated with control RNAi or *vg* RNAi. **i**, Comparing slopes of control RNAi ($n = 23$) and *vg* RNAi ($n = 35$); analysis of covariance (ANCOVA), $F = 38.1$, degrees of freedom (d.f.) $= 54$, $P < 0.0001$. Experiments were repeated at least three times. **j**, Wild-type minor worker and treated individuals with either rudimentary forewing disc or leg disc ablated. Red asterisk, ablated leg. **k**, Comparing slopes of leg disc ($n = 16$) and rudimentary forewing disc ablations ($n = 16$); ANCOVA, $F = 8.74$, d.f. $= 28$, $P = 0.0063$. Image comparisons are to scale. Experiments were repeated at least twice. All regressions are x axis (log(body length (µm)) versus y axis (log(head width (µm)).

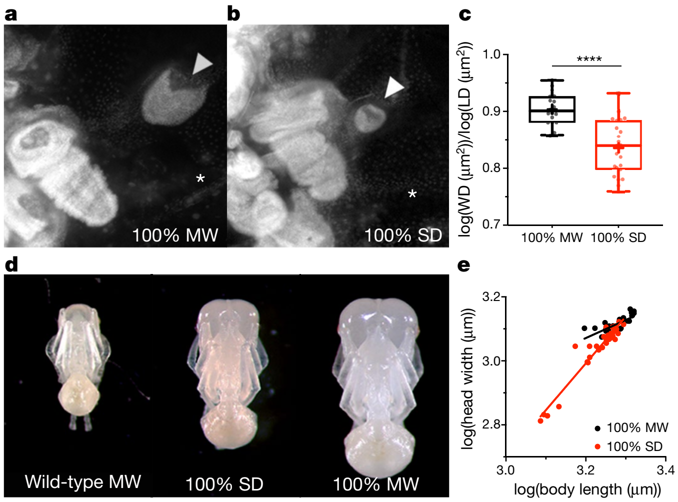


Fig. 3 | Social inhibition regulates size and disproportionate head-to-body scaling by suppressing growth of rudimentary forewing discs. Manipulations are on soldier-destined larvae. Rudimentary wing disc presence (arrowheads) or absence (asterisks). **a**, **b**, Comparing rudimentary wing disc size after no inhibition (100% minor workers) (**a**) and high inhibition (100% soldiers) (**b**). **c**, Comparing ratios of log(rudimentary forewing disc area (μm^2)) to log(leg disc area (μm^2)) between 100% minor workers ($n = 18$) and 100% soldiers ($n = 23$). The box plot shows mean (+), interquartile range (bars) and minimum to maximum values (whiskers); all points represent individual ants. Two-tailed unequal variance t-test, $t = 5.77$, d.f. = 36.13, ****$P < 0.0001$. **d**, Wild-type minor worker and representative individuals raised by 100% minor workers or 100% soldiers. **e**, Comparing slopes of 100% minor workers ($n = 24$) and 100% soldiers ($n = 35$); ANCOVA, $F = 36.55$, d.f. = 55, $P < 0.0001$; x axis (log(body length (μm)) versus y axis (log(head width (μm)). Image comparisons are to scale. Experiments were repeated at least three times.

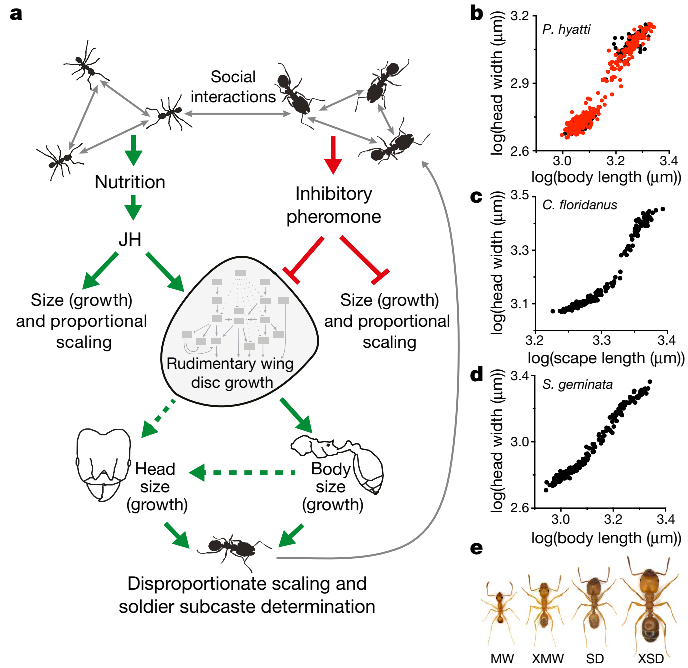


Fig. 4 | The role of rudimentary wing discs in the social regulation, development and evolution of complex worker-caste systems in ants. **a**, Interactions (arrows and lines) may be direct or indirect. Green arrows, activation; dashed green arrows, potential pathways to disproportionate scaling; red arrows and lines, inhibition; grey arrows, social interactions; grey circle, rudimentary forewing discs; grey boxes, wing gene network. **b**, Experimental manipulations (red) and controls (black) in *P. hyatti*. **c**, Wild-type *C. floridanus* workers[29] ($n = 179$). **d**, Wild-type *S. geminata* workers[30] ($n = 239$). Plots in **b**–**d** show log(body length (μm)) or log(scape length (μm)) on the x axis, versus log(head width (μm)) on the y axis. **f**, Comparison, to scale, of wild-type minor worker, large minor worker anomaly (XMW), wild-type soldier and supersoldier-like anomaly (XSD).

results similar to the larvae raised entirely by soldiers (Extended Data Fig. 7l–p). Taken together, the soldier inhibitory pheromone regulates size and disproportionate head-to-body scaling by suppressing growth of rudimentary forewing discs.

We next investigated whether rudimentary forewing discs mediate the interaction between the inhibitory pheromone and juvenile hormone to regulate size, head-to-body scaling and worker subcaste determination. Previous work[24] has demonstrated that the inhibitory pheromone acts downstream of juvenile hormone and can inhibit activation of soldier development—bipotential larvae treated with juvenile hormone and raised entirely by soldiers can block soldier development, which results in the production of large minor workers[24]. In *P. hyatti*, we found that these large minor workers develop from larvae that lack rudimentary forewing discs and are proportionally larger; although the intercept is significantly different relative to controls, the head-to-body slope is not (Extended Data Fig. 8i–l). Conversely, treating bipotential larvae with juvenile hormone but without inhibitory pheromone (100% minor workers) produced soldiers that crossed the juvenile-hormone threshold, as well as large minor workers that did not (Extended Data Fig. 8a–d). These large minor workers, which develop from larvae with rudimentary forewing discs that prematurely stop growing, are disproportionately larger; the head-to-body slope is significantly different compared to controls (Extended Data Fig. 8e–h). These results show that: (1) blocking growth of rudimentary forewing discs with inhibitory pheromone blocks the disproportionate head-to-body scaling that is induced by juvenile hormone; and (2) juvenile hormone regulates size and proportional head-to-body scaling independently of rudimentary forewing discs (Fig. 4a). Finally, to determine the role of inhibitory pheromone independent of rudimentary forewing discs, we exposed bipotential larvae to inhibitory pheromone (100% soldiers) without treatment with juvenile hormone. The minor workers that result develop from larvae that lack rudimentary forewing discs and are proportionally smaller; although the intercept is significantly different, the head-to-body slope is not (Extended Data Fig. 8m–p). Taken together, we show that juvenile hormone and the inhibitory pheromone regulate size and proportional head-to-body scaling independently of rudimentary forewing discs (Fig. 4a). By contrast, juvenile hormone and the inhibitory pheromone regulate size, disproportionate head-to-body scaling, and soldier subcaste determination, by directly or indirectly influencing the growth of rudimentary forewing discs (Fig. 4a).

Multiple molecular mechanisms may underlie the function of rudimentary wing discs. Organs, including wing imaginal discs in insects, communicate through conserved signalling pathways to coordinate development[27]. In flies, butterflies and beetles, eliminating or reducing the growth of one imaginal disc either increases the size of others owing to resource competition, or has no adult phenotypic effect owing to homeostasis[28]. By contrast, we show that eliminating or reducing the growth of a rudimentary wing disc decreases the head/body ratio. The rudimentary forewing discs may increase head growth secondarily, through their influence on body growth. Alternatively, because rudimentary forewing discs increase head size more than body size, they may increase head and body growth independently of each other (Fig. 4a, dashed lines). Therefore, this ancient communication system may have been exploited and modified in *Pheidole*, enabling rudimentary wing discs to regulate head-to-body allometry at the level of the individual and the subcaste ratio at the colony level in response to environmental variation and challenges[22–25] (Fig. 4a).

Wilson proposed[3] that worker-caste systems in ants—which range from being monomorphic to completely dimorphic—have evolved through a series of developmental transitions in allometry. Our experimental manipulations of the growth of rudimentary forewing discs

produced intermediate variants that fill the empty phenotypic space between minor workers and soldiers (compare Fig. 1c to Fig. 4b). These manipulations transform *P. hyatti* from a completely dimorphic worker-caste system to a sigmoid-like allometry, which mimics the worker-caste system of other species such as *Camponotus floridanus*[29] or *Solenopsis geminata*[30] (Fig. 4b–d). Furthermore, we discovered rare and anomalously large minor workers in a wild *P. hyatti* colony (Fig. 4e), which suggests that the intermediates we produced experimentally can appear in nature but that selection for a robust juvenile-hormone-mediated threshold between minor workers and soldiers may limit their production. We also discovered rare supersoldier-like anomalies in a wild *P. hyatti* colony (Fig. 4e). Because an ancestral developmental potential to produce supersoldiers has been retained across *Pheidole*, such anomalies can be experimentally induced with juvenile hormone[10]. Supersoldier anomalies develop from larvae with two pairs of rudimentary wing discs[10], which mimics the development of supersoldiers in *P. obtusospinosa*[10] (Fig. 1f, i). These findings show that rudimentary wing discs store an ancestral developmental potential that—when released—produces allometric variation that mimics the worker-caste systems of other species (Fig. 4b–d). This potential may facilitate 'Wilson's transitions'[3] between different complex worker-caste systems, and their parallel evolution across the ants (Extended Data Fig. 1).

The transient appearance of rudimentary organs is a general feature of organismal development, and is primarily used to infer common ancestry[1,12–14]. Although the functional importance of rudimentary organs is enigmatic[1,12–14], we show that rudimentary wing discs have a major regulatory role during ant development and evolution. We propose that rudimentary organs may generally evolve key regulatory functions and, through their capacity to store and release ancestral developmental potential, may be an underappreciated source of variation that fuels adaptive evolution.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0613-1.

Received: 7 December 2017; Accepted: 22 August 2018;
Published online 10 October 2018.

1. Darwin, C. *On the Origin of Species* (John Murray, London, 1859).
2. Huxley, J. *Problems of Relative Growth* (Metheun, London, 1932).
3. Wilson, E. O. The origin and evolution of polymorphism in ants. *Q. Rev. Biol.* **28**, 136–156 (1953).
4. Wheeler, D. E. The developmental basis of worker caste polymorphisms in ants. *Am. Nat.* **138**, 1218–1238 (1991).
5. Molet, M., Wheeler, D. E. & Peeters, C. Evolution of novel mosaic castes in ants: modularity, phenotypic plasticity, and colonial buffering. *Am. Nat.* **180**, 328–341 (2012).
6. Metzl, C., Wheeler, D. E. & Abouheif, E. Wilhelm Goetsch (1887–1960): pioneering studies on the development and evolution of the soldier caste in social insects. *Myrmecol. News* **26**, 81–96 (2018).
7. Abouheif, E. & Wray, G. A. Evolution of the gene network underlying wing polyphenism in ants. *Science* **297**, 249–252 (2002).
8. Wilson, E. O. *Pheidole in the New World* (Harvard Univ. Press, Cambridge, 2003).
9. Wheeler, D. E. & Nijhout, H. F. Imaginal wing discs in larvae of the soldier caste of *Pheidole bicarinata vinelandica* Forel (Hymenoptera: Formicidae). *Int. J. Insect Morphol. Embryol.* **10**, 131–139 (1981).
10. Rajakumar, R. et al. Ancestral developmental potential facilitates parallel evolution in ants. *Science* **335**, 79–82 (2012).
11. Sameshima, S.-Y., Miura, T. & Matsumoto, T. Wing disc development during caste differentiation in the ant *Pheidole megacephala* (Hymenoptera: Formicidae). *Evol. Dev.* **6**, 336–341 (2004).
12. Gould, S. J. *Ontogeny and Phylogeny* (Belknap, Cambridge, 1977).
13. Mayr, E. Recapitulation reinterpreted: the somatic program. *Q. Rev. Biol.* **69**, 223–232 (1994).
14. Hall, B. K. Descent with modification: the unity underlying homology and homoplasy as seen through an analysis of development and evolution. *Biol. Rev. Camb. Philos. Soc.* **78**, 409–433 (2003).
15. Economo, E. P. et al. Global phylogenetic structure of the hyperdiverse ant genus *Pheidole* reveals the repeated evolution of macroecological patterns. *Proc. R. Soc. Lond. B* **282**, 20141416 (2015).
16. Wheeler, D. E. & Nijhout, H. F. Soldier determination in ants: new role for juvenile hormone. *Science* **213**, 361–363 (1981).
17. Pagel, M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* **255**, 37–45 (1994).
18. Williams, J. A., Bell, J. B. & Carroll, S. B. Control of *Drosophila* wing and haltere development by the nuclear vestigial gene product. *Genes Dev.* **5**, 2481–2495 (1991).
19. Van de Bor, V., Delanoue, R., Cossard, R. & Silber, J. Truncated products of the *vestigial* proliferation gene induce apoptosis. *Cell Death Differ.* **6**, 557–564 (1999).
20. Kim, J. et al. Integration of positional signals and regulation of wing formation and identity by *Drosophila* vestigial gene. *Nature* **382**, 133–138 (1996).
21. Zider, A., Flagiello, D., Frouin, I. & Silber, J. *Vestigial* gene expression in *Drosophila melanogaster* is modulated by the dTMP pool. *Mol. Gen. Genet.* **251**, 91–98 (1996).
22. Yang, A. S., Martin, C. H. & Nijhout, H. F. Geographic variation of caste structure among ant populations. *Curr. Biol.* **14**, 514–519 (2004).
23. Lillico-Ouachour, A. & Abouheif, E. Regulation, development, and evolution of caste ratios in the hyperdiverse ant genus *Pheidole*. *Curr. Opin. Insect Sci.* **19**, 43–51 (2017).
24. Wheeler, D. E. & Nijhout, H. F. Soldier determination in *Pheidole bicarinata*: inhibition by adult soldiers. *J. Insect Physiol.* **30**, 127–135 (1984).
25. Passera, L., Roncin, E., Kaufmann, B. & Keller, L. Increased soldier production in ant colonies exposed to intraspecific competition. *Nature* **379**, 630–631 (1996).
26. Lillico-Ouachour, A. *The Behavioural, Chemical, and Morphological Basis of Caste Regulation in the Worker Caste of Ants*. MSc thesis, McGill Univ. (2017).
27. Droujinine, I. A. & Perrimon, N. Interorgan communication pathways in physiology: focus on *Drosophila*. *Annu. Rev. Genet.* **50**, 539–570 (2016).
28. Shingleton, A. W. & Frankino, W. A. The (ongoing) problem of relative growth. *Curr. Opin. Insect Sci.* **25**, 9–19 (2018).
29. Alvarado, S., Rajakumar, R., Abouheif, E. & Szyf, M. Epigenetic variation in the *Egfr* gene generates quantitative variation in a complex trait in ants. *Nat. Commun.* **6**, 6513 (2015).
30. Tschinkel, W. R. The morphometry of *Solenopsis* fire ants. *PLoS ONE* **8**, e79559 (2013).

Acknowledgements We thank R. Johnson, K. Haight, A. Wild, L. Davis, R. Sanwald and A. Nisip for help collecting ants; J. Liebig for help with cuticular hydrocarbon experiments; T. Oakley and P. Ward for help with phylogenetic analyses; and Y. Tomoyasu, I. Ruvinsky, B. Hall, D. E. Wheeler, D. Schoen, A. Shingleton, V. Callier, G. Wray, S. C. Weber, members of the Abouheif Laboratory, M. J. West-Eberhard and E. O. Wilson for comments on the manuscript. We thank the McGill University Advanced BioImaging Facility for imaging support. This work was supported by KLI fellowships (Austria) to R.R. and E.A., and NSERC Discovery Grant and Steacie Fellowship (Canada) and Guggenheim Fellowship (USA) to E.A.

Author contributions E.A. and R.R. conceived the project and designed experiments; E.A., R.R., A.R. and S.K. collected ants; R.R., S.K., M.C. and A.R. performed in situ hybridization and immunohistochemistry; R.R., M.-J.F., M.C., S.K. and T.C. performed RNAi; S.K. performed ablations; R.R., G.D.B., A.L.-O., M.C. and T.C. performed pheromone experiments; R.R., M.C., M.-J.F., A.L.-O., S.K., G.D.B. and T.C. performed hormone experiments; T.C. and R.R. performed semi-quantitative PCR; E.A. performed phylogenetic analyses; and D.O. mounted adult specimens. E.A. and R.R. wrote the manuscript with input from co-authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0613-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0613-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to E.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Animal collection and culturing. Colonies of *P. hyatti* and *P. obtusospinosa* were collected in the southwest of Arizona, USA. *C. floridanus* colonies were collected as newly mated queens in Tallahassee, Florida, USA. *Monomorium trageri* colonies and *S. geminata* larvae were collected in Gainesville, Florida, USA. All colonies and experimental replicates were kept at 27 °C with 60% humidity and a 12:12-h light:dark cycle. They were kept in fluon-coated plastic boxes with cotton-constrained glass water tubes as well as sugar water tubes. They were fed mealworms and Bhatkar–Whitcomb diet[31]. Holes were inserted in the lids of the boxes and lined with mesh to allow better air exchange. All experiments were performed in accordance with invertebrate care guidelines and regulations.

Testing for phylogenetic correlation. Pagel's[17] maximum likelihood based phylogenetic correlation test, as implemented in Mesquite 3.5[32], was used to test for a phylogenetic correlation between the evolution of inter-subcaste variation in size of rudimentary wing discs and evolution of a soldier subcaste. One hundred extra iterations were used to set the intensity of the likelihood search. Pagel's[17] method first calculates the log-likelihood that the presence of inter-subcaste variation in the size of rudimentary wing discs and the presence of soldier subcastes evolved independently of each other along the phylogeny, and then calculates the log-likelihood that both traits evolved in a correlated manner along the phylogeny. It then uses a likelihood ratio test statistic (1,000 simulations of the data were used) to determine which of these two models best fits the data. For data on the size of rudimentary wing discs and soldier subcastes, all available comparative data from the literature and from the present study were collected, resulting in data points for 21 species across 10 ant genera (Extended Data Fig. 1 and Extended Data Table 1). Phylogenetic relationships and branch-length information were obtained for all 21 species (Extended Data Fig. 1a) from previous publications[33–40], and provide formal phylogenetic evidence that a soldier subcaste has evolved several times independently across the ants. For each species, the presence of inter-subcaste variation in size of rudimentary wing discs was coded as a '1' if rudimentary wing discs were discretely larger in the soldier subcaste compared to the minor-worker subcaste (Extended Data Table 1). Inter-subcaste variation in size of rudimentary wing discs was coded as '0' if inter-subcaste variation was absent and larvae in the colony had similar sizes of rudimentary wing discs (Extended Data Table 1). For each species, the presence of a soldier subcaste was coded as '1' if it was demonstrated by previous studies (Extended Data Table 1). The absence of a soldier subcaste was coded as '0' if it was demonstrated by previous studies (Extended Data Table 1).

Isolation of *vestigial* homologues. Forward: 5′-TATCCTTACCTKTAYCARACCC-3′ and reverse: 5′-GCTGACTATTCCAAAAGGARGG-3′ degenerate primers were designed based on a *vestigial* sequence alignment of *Tribolium castaneum* (XM_008201106), *Apis mellifera* (XM_001122002.4) and several ant species obtained from the Ant Genome Portal database (http://hymenopteragenome.org/ant_genomes/). *P. hyatti* RNA was isolated using TRIzol (Invitrogen) from a pool of embryos and larvae of different developmental stages. RNA was then reverse-transcribed to synthesize a cDNA library. PCR amplicons were ligated into the pGEM-T easy vector (Promega) and subsequently sequenced (MH683613) using Sanger sequencing at the Genome Quebec Innovation Centre at McGill University. To determine whether these *vestigial* sequences—including the TONDU/vestigial domain—are conserved, amino acid alignment was performed using Geneious Alignment on Geneious (R8), followed by manual alignment (Supplementary Fig. 2).

Whole-mount in situ hybridization. Terminal stage embryos and larvae were fixed and processed as previously described[7,41–43]. Using a Zeiss Discovery V12 stereomicroscope, fixed samples were then dissected to remove the gut and fat tissue to facilitate probe access to imaginal tissues. Digoxigenin-labelled riboprobes (Roche Diagnostics Canada) for *vestigial* and *wingless*[7] were then synthesized from the cloned *P. hyatti* fragments and subsequently used for spatial gene expression profiling in embryos[42] and larvae[44]. Reproductive, soldier and minor worker larvae were pooled during in situ hybridization.

Semi-quantitative reverse-transcription PCR of *vestigial* transcripts in *P. hyatti*. *vestigial* expression in wild-type terminal *P. hyatti* larvae was determined through the detection of *vestigial* transcripts using semi-quantitative PCR with reverse transcription (Extended Data Fig. 3 and Supplementary Fig. 1). The housekeeping gene *Elongation factor 1 alpha* (*EF1a*) was used as our reference gene because it has previously been validated as a reference gene for quantitative PCR with reverse transcription in social insects[45]. Therefore, *EF1a* was first cloned and sequenced (MH683615) from a *P. hyatti* cDNA library using the following degenerate primers: forward 5′-GATTCYGGCAAGTCGACCA-3′ and reverse 5′-GGAACTCTTGGAAAGCCTCAAC-3′. The PCR product was sequenced using Sanger sequencing at the Genome Quebec Innovation Centre at McGill University. To extract RNA, three minor worker larvae and three soldier larvae at the terminal stage were collected from a laboratory colony and total RNA was extracted from minor workers and soldiers separately. The tissue was disrupted using a TissueLyser (Qiagen) bead mill and RNA was extracted using the TRIzol (Invitrogen) RNA extraction protocol[46], and then purified using the RNeasy Plus Kit (Qiagen). Minor worker and soldier RNA was treated with DNase I (Invitrogen) to remove genomic DNA before being reverse transcribed into cDNA using the Superscript III First-Strand Synthesis System (Invitrogen). The concentrations of total RNA and total cDNA were normalized between minor workers and soldiers before cDNA synthesis and PCR, respectively. The two cDNA libraries were used as PCR templates for the semi-quantitative PCR with reverse transcription of *vestigial* and *EF1a*. The *P. hyatti vestigial* PCR primers used were: forward 5′-TCCTTACCTGTATCAGACCCATC-3′ and reverse 5′-TGTCGATCTGTCGTCGTCCAA-3′, and the *P. hyatti EF1a* PCR primers used were: forward 5′-TCAGGACGTGTACAAGATC-3′ and reverse 5′-CAATGACCTGTGCAGTAAAG-3′. The PCR was performed using an annealing temperature of 56 °C with 31 thermocycles and four serial dilutions of the two cDNA libraries (500 ng, 250 ng, 125 ng and 62.5 ng) were used for the semi-quantitative PCR with reverse transcription. A no-template water control was used to confirm absence of contamination.

Phospho-histone H3 (PH3) assay. To assay for proliferation, rudimentary wing imaginal discs from soldier-destined larvae measuring 2.2 mm, 2.5 mm and 2.7 mm in size were fixed and dissected (as described in 'Whole-mount in situ hybridization'), and then immunohistochemistry against PH3 was performed as previously described[47]. Mouse monoclonal anti-PH3 (PH3 Ser10, Cell Signaling Technology, 9706S) primary antibody was used at a 1:25 dilution and the secondary antibody, goat anti-mouse Alexa 555 (AbCam, AB150114) was used at a 1:500 dilution. Samples were counter-stained with DAPI.

RNAi. To functionally knock down *vestigial*, *vg* double-stranded RNA (dsRNA) was synthesized from the same plasmid from which the *vg* in situ riboprobe was synthesized. The fragment (not including primer regions) was 631 bp, which is of a size range that is known to be highly efficient in targeting imaginal discs of other insects[48]. T7-flanked (uppercase) *vg* forward 5′-TAATACGACTCACTATAGGGtatccttacctgtaccagaccc-3′ and reverse 5′-TAATACGACTCACTATAGGGgctgactattccaaaaggaggg-3′ primers were used to amplify a template for T7 RNA polymerase to synthesize the dsRNA[49,50]. dsRNA was then purified with the Qiagen RNAeasy purification kit followed by ethanol precipitation and eluted in Spradling injection buffer[51] to a concentration of 3.5 mg/ml.

Size-matched soldier-destined larvae were then injected with *vestigial* dsRNA or with *yfp* dsRNA or injection buffer as a control. Soldier-destined larvae can be defined as larvae that ranged from 2.2 to 2.75 mm, and these larvae possess a characteristic brown gut[29]. Larvae at this stage were specifically injected, because this is when the rudimentary forewing imaginal discs appear and begin to grow in *Pheidole*[10] and at this stage *vg* expression in the wing discs of *Drosophila* is required[18]. To test whether *vg* RNAi affects head-to-body scaling independently of the wing disc, size-matched bipotential larvae—which are defined as ranging from 1.0 to 1.6 mm[10]—were injected with *vg* dsRNA because 95% of bipotential larvae develop into minor worker larvae that lack rudimentary wing discs. Finally, to determine the effect of *vg* RNAi on wing imaginal discs that develop into functional wings, size-matched male-destined larvae were injected, and wing phenotypes were examined and head width and body size (Weber's length) were measured.

To inject soldier-destined, bipotential and male larvae, the larvae were placed onto two-sided tape on a microscope slide and stabilized along a glass capillary tube. Needles were made using a Sutter Instrument needle puller (Model P-97) and inserted onto a Narishige microinjection apparatus attached to a Zeiss Discovery V8 dissection microscope fitted with a custom-made *x*–*y* movable platform. Microinjection flow pressure was controlled using a Cell Tram Vario 5176 (Eppendorf). Larvae were injected, while on their dorsum, through their lateral side at the midline of the antero-posterior axis. After the injection, larvae were placed in replicates containing adult minor workers to care for them (in a 2:1 adult:larva ratio) until they were either fixed before pupation and DAPI-stained for imaginal disc imaging and measurements, left to metamorphose into pupae for pupal imaging and measurements or—in the case of males—left to become adults for measurements.

Efficiency and specificity of *vg* RNAi. Semi-quantitative reverse-transcription PCR was performed on terminal *P. hyatti* soldier larvae after injection with *vg* RNAi or *yfp* RNAi to confirm knockdown of *vg* transcripts (Extended Data Fig. 4 and Supplementary Fig. 1). RNA was extracted from terminal soldier larvae 48 h post-injection and cDNA libraries were synthesized and standardized as described above. *NADH* (*NADH-ubiquinone oxidoreductase subunit 8*) was used as a reference gene as it has previously been validated as a reference gene for quantitative PCR in ants[52]. Therefore, *NADH* was first cloned and sequenced (MH683614) from a *P. hyatti* cDNA library using the following degenerate primers: forward 5′-GGGBCCTTTACAGATAATTGCRC-3′ and reverse 5′-ATTCRAAGTTRGGACCCTCA-3′. The two cDNA libraries (*vg* RNAi and *yfp* RNAi) were used as PCR templates for the semi-quantitative reverse-transcription PCR of *vestigial* and *NADH*. The *NADH P. hyatti*

primers were: forward 5′-GGGAGAGCATGCGTTAAGAA-3′ and reverse 5′-TAGCCTGTGCAGGACAAATC-3′. The *vestigial* primers targeting outside the dsRNA fragment were: forward 5′-CATTACCCACAGTACCATCACA-3′ and reverse 5′-CAGCAGAAGGCCACTGTAG-3′. The PCR was performed using an annealing temperature of 60 °C with 35 thermocycles and four serial dilutions of the two cDNA libraries were used as template. A no-template water control was used to confirm absence of contamination.

Finally, the TUNEL (TdT-mediated dUTP nick end labelling) assay was used to validate the specificity of *vg* RNAi because perturbation of *vg* expression is known to induce apoptosis in *Drosophila* wing imaginal discs[19]. Apoptotic cell death was assayed using the In Situ Cell Death Detection Kit, AP (Roche Diagnostics). The protocol was carried out as in a previous publication[53], with a modification of the proteinase K step to 5 min (50 μg/ml).

Ablations. The ablation protocol was modified from a previous publication[54]. Soldier-destined larvae were collected, as described in 'RNAi'. Leg and rudimentary wing disc ablations were done using a Hyfrecator 2000 electrosurgical unit (ConMed) with an electrochemically sharpened (1 M NaOH and 120 V) tungsten wire. The tungsten wire was fastened and wound around an extra-fine needle electrode (714, ConMed) to enable current to pass through the sharpened wire. All ablations were performed using 1.5 W of electrical current for 0.5–1 s. During cauterization, larvae were placed on their dorsum on a KimWipe moistened with PBS under a Zeiss Discovery V8 dissection microscope to enable visualization of the leg discs. The left rudimentary wing disc was cauterized using the position of the leg discs and boundaries of the thoracic segments as landmarks. To control for the effect of cauterization and for damage to imaginal tissue, the left posterior leg imaginal disc was cauterized using the same intensity and time of cauterization. After cauterization, larvae were placed in replicates containing adult minor workers to care for them (in a 2:1 adult:larva ratio) until they were either fixed before pupation and DAPI-stained for imaginal disc imaging or left to metamorphose into pupae for pupal imaging and measurements. Analysing the effect of control leg disc ablations on pupae and dissected larvae allowed us to confirm that cauterization was specific to the targeted tissue and did not damage surrounding imaginal tissues (Fig. 2j and Extended Data Fig. 5m–q).

Soldier inhibition and juvenile hormone manipulations. Previous studies have shown that exposure of *Pheidole* larvae to a high density of adult soldiers reduces their potential to develop into soldiers owing to the production of a soldier inhibitory pheromone[23]. In particular, bipotential larvae that were treated with a dose of the juvenile hormone analogue methoprene—which would normally induce soldier development—became minor workers when raised by 100% soldiers[24]. In the absence of adult soldiers, treating size-matched bipotential larvae with methoprene has previously been shown to induce soldier and supersoldier development in *Pheidole* and results in the development of rudimentary wing discs[10]. To determine whether the inhibition of soldier development is mediated by rudimentary wing disc growth, sized-matched bipotential larvae were treated with methoprene (5 mg/ml, Sigma-Aldrich) or acetone vehicle control and raised by either 100% minor workers (no inhibition) or 100% soldiers (high inhibition) in a 2:1 adult:larva ratio. Larvae were either fixed before pupation for imaginal disc imaging or left to metamorphose into pupae for pupal imaging and measurements. To determine the effects of this inhibitory pheromone on soldier development, soldier-destined larvae were collected as described in 'RNAi', and placed in replicates that contained 100% adult soldiers (high inhibition) or 100% minor workers (no inhibition) in a 2:1 adult:larva ratio. Larvae were fixed before pupation and DAPI-stained for imaginal disc measurements or left to metamorphose into pupae for pupal imaging and measurements. To determine whether the inhibitory pheromone is involved in regulating head-to-body scaling independent of the disc, bipotential larvae—which do not possess rudimentary discs—were raised with 100% adult soldiers (in a 2:1 adult:larva ratio). These were additionally treated with acetone, to enable comparison to the acetone + 100% minor workers treatment described above. Individual ants were then either fixed as larvae for imaginal disc measurements or left to metamorphose for pupal measurements.

Cuticular hydrocarbon extraction and application. To test whether the soldier inhibitory pheromone is capable of changing head-to-body scaling independent of rearing environment, soldier-destined larvae were treated with cuticular hydrocarbons extracted in a 2:1 adult:larva ratio from adult soldiers. The extraction and application protocol was modified from a previous publication[55]. To extract the surface cuticular hydrocarbons from adult soldiers, adult soldiers were frozen at −80 °C for 3 min and then placed in a Teflon-capped borosilicate glass vial (Sigma-Aldrich) containing 50 μl hexane per 10 soldiers for 2 min. Extracts were dried with high-purity nitrogen and resuspended in hexane. Soldier-destined larvae were treated with 1 μl of extract or hexane control and then placed in replicates containing adult minor workers to care for them (in a 2:1 adult:larva ratio). Larvae were left to metamorphose into pupae for pupal imaging and measurements.

Microscopy and morphometrics. Larval and pupal imaging was conducted with a Zeiss Discovery V12 stereomicroscope, while the imaging of imaginal discs (Bright Field, DIC and fluorescence) was performed with a Zeiss AxioImager Z1 microscope. Larval rudimentary wing discs stained for PH3 were imaged with a confocal Leica SP8 Point-Standing Confocal system on a Leica DMI6000B Inverted Microscope. Zeiss AxioVision software was used to take larval, imaginal disc and pupal measurements. All rudimentary wing disc size measurements were performed on terminal stage larvae just before pupation, as this establishes a standardized stage for comparisons between castes and experimental manipulations[7]. Furthermore, rudimentary wing disc size was compared to leg disc size to normalize for variation among individual ants within the terminal stage. Head width and body length of pupae was measured in a dorsal view because insects obtain their final adult shape and size after they metamorphose, and measurements on pupae are more repeatable and accurate than those on adults.

Statistics and reproducibility. All statistical analyses were performed on Graphpad Prism v.7 and statistical analyses were considered statistically significant at a P value < 0.05. All measures were log-transformed before analysis, with the exception of those used to calculate percentage change in head width and body length. To estimate measurement error, 20 wild-type minor worker and 20 wild-type soldier pupae were randomly selected and pupal head width and body length were measured. Each morphological trait was measured three times and the person taking the measures was blind to previous values. A random-effects ANOVA (RStudio, v.1.1.423) was used to estimate the measurement error in log-transformed values for head width and body length. The measurement error was found to make up a negligible percentage of total variance in head width (0.0599%) and body length (0.0822%). Therefore, ordinary least squares linear regression was used throughout because previous work[56] has shown that the ordinary least squares method is more accurate than reduced major-axis regression when there is low measurement error in the variable on the x axis. No statistical methods were used to predetermine sample size, all experiments were independently repeated multiple times and converged on similar results as shown in the Figs. and Extended Data Figs. Blinding and randomization was used during all experimental manipulation.

To test whether any of our manipulations (*vg* RNAi, ablations, high inhibition raised by 100% soldiers, soldier CHC extracts and/or juvenile hormone treatments) altered head-to-body scaling compared to their controls, ANCOVA was used to test for differences in slope and—if no change in slope was observed—differences in y intercept between linear regressions of treatment and the respective control were assessed. To compare the ratio of rudimentary wing disc/leg disc area, or the ratio of pupal head width/body length, head width, or body length, normality was first tested using the Shapiro–Wilk test. If normality was violated, a non-parametric Mann–Whitney U-test was used, and if not an unpaired t-test was used. Where necessary, a Welch's correction for unequal variances was used. All pairwise comparisons were two-tailed except for (1) electrosurgical ablations; (2) percentage change in head width versus body length; and (3) cuticular hydrocarbon treatments, for which we used one-tailed tests based on a priori expectations derived from the results of our preliminary RNAi and soldier inhibition experiments. This workflow was used to determine whether *vg* RNAi, ablations, social and juvenile hormone manipulations, and CHC treatments differed compared to their respective controls for: (1) the ratio of average rudimentary wing disc area to average leg disc area; (2) the ratio of head width to body length; (3) head width; and (4) body length. For multiple pairwise comparisons of head width or body length between social and/or juvenile hormone manipulations and their controls, a Bonferroni correction was applied. Finally, Fisher's exact test was used to determine whether the proportion of males with affected wing morphology differed between *yfp* RNAi control males and *vg* RNAi.

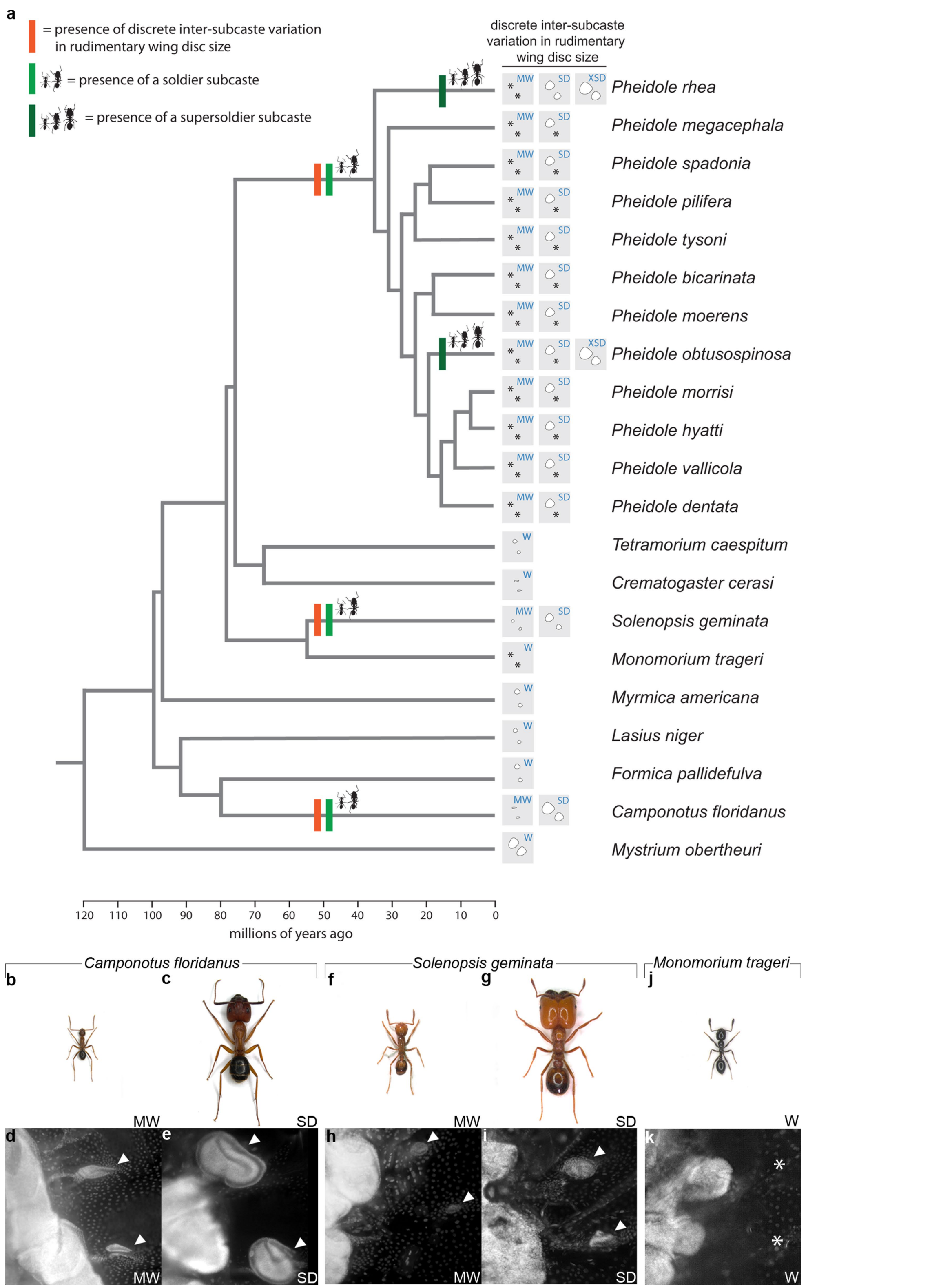
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

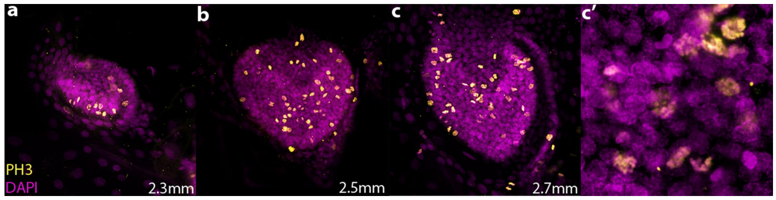
All relevant data are included in the paper. The raw data for all analyses used in this study are available from the corresponding author upon request.

31. Bhatkar, A. & Whitcomb, W. H. Artificial diet for rearing various species of ants. *Fla. Entomol.* **53**, 229–232 (1970).
32. Maddison, W. P. & Maddison, D. R. Mesquite: a modular system for evolutionary analysis. https://www.mesquiteproject.org/ (2018).
33. Moreau, C. S. & Bell, C. D. Testing the museum versus cradle tropical biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the ants. *Evolution* **67**, 2240–2257 (2013).
34. Ward, P. S., Brady, S. G., Fisher, B. L. & Schultz, T. R. The evolution of myrmicine ants: phylogeny and biogeography of a hyperdiverse ant clade (Hymenoptera: Formicidae). *Syst. Entomol.* **40**, 61–81 (2015).
35. Ward, P. S. The phylogeny and evolution of ants. *Annu. Rev. Ecol. Evol. Syst.* **45**, 23–43 (2014).
36. Ward, P. S., Blaimer, B. B. & Fisher, B. L. A revised phylogenetic classification of the ant subfamily Formicinae (Hymenoptera: Formicidae), with resurrection of the genera *Colobopsis* and *Dinomyrmex*. *Zootaxa* **4072**, 343–357 (2016).

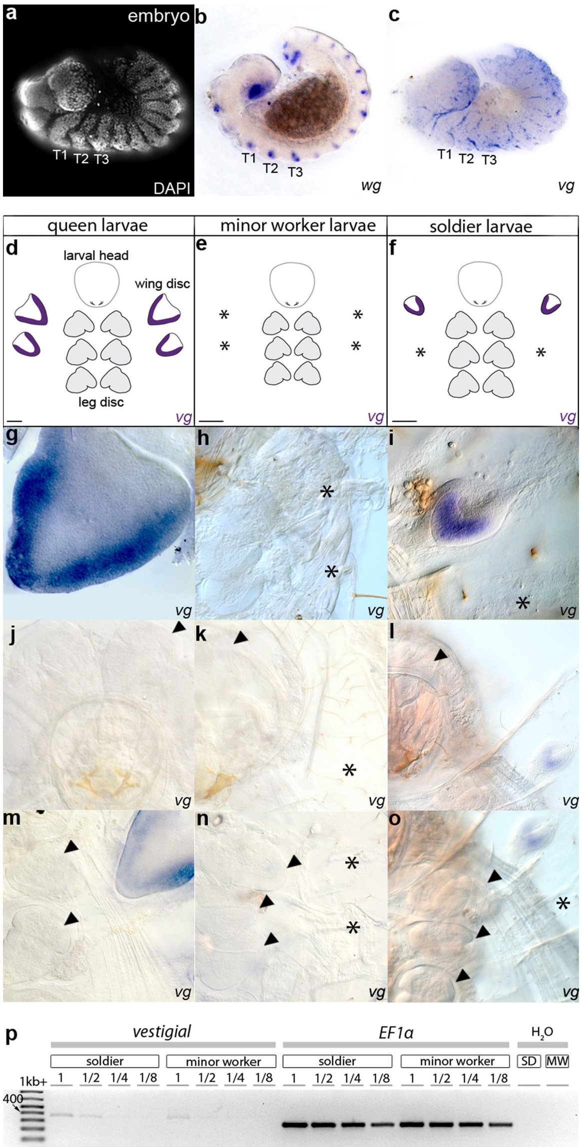
37. Blaimer, B. B. et al. Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC Evol. Biol*. **15**, 271 (2015).
38. Branstetter, M. G. et al. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. *Curr. Biol*. **27**, 1019–1025 (2017).
39. Moreau, C. S. Unraveling the evolutionary history of the hyperdiverse ant genus *Pheidole*. *Mol. Phylogenet. Evol*. **48**, 224–239 (2008).
40. Blanchard, B. D. & Moreau, C. S. Defensive traits exhibit an evolutionary trade-off and drive diversification in ants. *Evolution* **71**, 315–328 (2017).
41. Patel, N. H. Imaging neuronal subsets and other cell types in whole-mount *Drosophila* embryos and larvae using antibody probes. *Methods Cell Biol*. **44**, 445–487 (1994).
42. Khila, A. & Abouheif, E. In situ hybridization on ant ovaries and embryos. *Cold Spring Harb. Protoc*. **2009**, doi:10.1101/pdb.prot5250 (2009).
43. Wheeler, G. C. & Wheeler, J. *Ant Larvae: Review and Synthesis* Memoirs of the Entomological Society of Washington Vol. 7 (The Entomological Society of Washington, Gainesville, 1976).
44. Tautz, D. & Pfeifle, C. A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene hunchback. *Chromosoma* **98**, 81–85 (1989).
45. Lourenço, A. P., Mackert, A., dos Santos Cristino, A. & Simões, Z. L. P. Validation of reference genes for gene expression studies in the honey bee, *Apis mellifera*, by quantitative real-time RT–PCR. *Apidologie (Celle)* **39**, 372–385 (2008).
46. Khila, A., Abouheif, E. & Rowe, L. Function, developmental genetics, and fitness consequences of a sexually antagonistic trait. *Science* **336**, 585–589 (2012).
47. Favé, M. J. et al. Past climate change on Sky Islands drives novelty in a core developmental gene network and its phenotype. *BMC Evol. Biol*. **15**, 183 (2015).
48. Miller, S. C., Miyata, K., Brown, S. J. & Tomoyasu, Y. Dissecting systemic RNA interference in the red flour beetle *Tribolium castaneum*: parameters affecting the efficiency of RNAi. *PLoS ONE* **7**, e47431 (2012).
49. Khila, A., Abouheif, E. & Rowe, L. Evolution of a novel appendage ground plan in water striders is driven by changes in the *Hox* gene *Ultrabithorax*. *PLoS Genet*. **5**, e1000583 (2009).
50. Khila, A., Abouheif, E. & Rowe, L. Comparative functional analyses of ultrabithorax reveal multiple steps and paths to diversification of legs in the adaptive radiation of semi-aquatic insects. *Evolution* **68**, 2159–2170 (2014).
51. Spradling, A. C. & Rubin, G. M. Transposition of cloned P elements into *Drosophila* germ line chromosomes. *Science* **218**, 341–347 (1982).
52. Livramento, K. G. D., Freitas, N. C., Máximo, W. P. F., Zanetti, R. & Paiva, L. V. Gene expression profile analysis is directly affected by the selected reference gene: the case of leaf-cutting *Atta sexdens*. *Insects* **9**, 18 (2018).
53. Shbailat, S. J., Khila, A. & Abouheif, E. Correlations between spatiotemporal changes in gene expression and apoptosis underlie wing polyphenism in the ant *Pheidole morrisi*. *Evol. Dev*. **12**, 580–591 (2010).
54. Nijhout, H. F. Pattern formation on lepidopteran wings: determination of an eyespot. *Dev. Biol*. **80**, 267–274 (1980).
55. Penick, C. A. & Liebig, J. A larval 'princess pheromone' identifies future ant queens based on their juvenile hormone content. *Anim. Behav*. **128**, 33–40 (2017).
56. Kilmer, J. T. & Rodríguez, R. L. Ordinary least squares regression is indicated for studies of allometry. *J. Evol. Biol*. **30**, 4–12 (2017).
57. Huang, M. H. & Wheeler, D. E. Colony demographics of rare soldier-polymorphic worker caste systems in *Pheidole* ants (Hymenoptera, Formicidae). *Insect. Soc*. **58**, 539–549 (2011).
58. Shbailat, S. J. & Abouheif, E. The wing-patterning network in the wingless castes of Myrmicine and Formicine ant species is a mix of evolutionarily labile and non-labile genes. *J. Exp. Zool. B Mol. Dev. Evol*. **320**, 74–83 (2013).
59. Bharti, H. & Kumar, R. Taxonomic studies on genus *Tetramorium* Mayr (Hymenoptera, Formicidae) with report of two new species and three new records including a tramp species from India with a revised key. *ZooKeys* **207**, 11–35 (2012).
60. Bolton, B. *Synopsis and Classification of Formicidae* Memoirs of the American Entomological Institute Vol. 71 (American Entomological Institute, Gainesville, 2003).
61. Morgan, C. E. & Mackay, W. *The North America Acrobat Ants of the Hyperdiverse Genus* Crematogaster *(Hymenoptera: Formicidae)* (Lambert, Saarbrücken, 2017).
62. DuBois, M. B. A revision of the native New World species of the ant genus *Monomorium* (minimum group) (Hymenoptera: Formicidae). *Univ. Kans. Sci. Bull*. **53**, 65–119 (1986).
63. Wilson, E. O. Division of labor in fire ants based on physical castes (Hymenoptera: Formicidae: Solenopsis). *J. Kans. Entomol. Soc*. **51**, 615–636 (1978).
64. Weber, N. A. Description of new North American species and subspecies of Myrmica Latreille (Hym.: Formicidae). *Lloydia* **2**, 144–152 (1939).
65. Wilson, E. O. A monographic revision of the ant genus *Lasius*. *Bull. Mus. Comp. Zool*. **113**, 1–204 (1955).
66. Trager, J. C., MacGown, J. A. & Trager, M. D. in *Advances in Ant Systematics (Hymenoptera: Formicidae): Homage to E. O. Wilson — 50 Years Of Contributions* Memoirs of the American Entomological Institute Vol. 80 610–636 (American Entomological Institute, Gainesville, 2007).
67. Béhague, J. et al. Lack of interruption of the gene network underlying wing polyphenism in an early-branching ant genus. *J. Exp. Zool. B Mol. Dev. Evol*. **330**, 109–117 (2018).



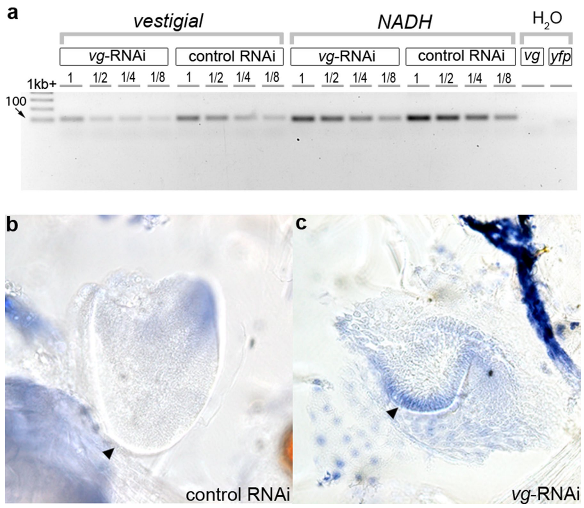
Extended Data Fig. 1 | The phylogenetic correlation between presence of a soldier subcaste and presence of discrete inter-subcaste variation in rudimentary wing disc size. a, Pagel's test (difference in log likelihoods = 8.43, $P = 0.001$) shows that there is a significant phylogenetic correlation between the presence of a soldier subcaste and the presence of discrete inter-subcaste variation in the size of rudimentary wing discs, for 21 species of ants. Grey lines indicate phylogenetic relationships, and the time scale of these relationships is indicated at the bottom in millions of years. The orange bar indicates presence of discrete inter-subcaste variation in the size of rudimentary wing discs, the green bar with two adjacent cartoons indicates presence of minor worker and soldier subcastes, and the dark green bar with three adjacent cartoons indicates presence of minor worker, soldier and supersoldier subcastes. Inter-subcaste variation in the size of rudimentary wing discs in workers (W) and/or minor workers and/or soldiers and/or supersoldiers are at the tip of each branch, where the white-circle drawings represent presence and relative size, within species, of rudimentary wing discs and asterisks indicate absence of rudimentary wing discs. Extended Data Table 1 provides references and descriptions. **b**–**k**, Adults and rudimentary wing discs of minor workers and/or soldiers and/or workers of *C. floridanus* (**b**–**e**), *S. geminata* (**f**–**i**) and *M. trageri* (**j**, **k**). **d**, **e**, **h**, **i**, **k**, Arrowheads indicate the presence of rudimentary wing discs and asterisks indicate the absence of rudimentary wing discs. All comparisons within species are to scale.



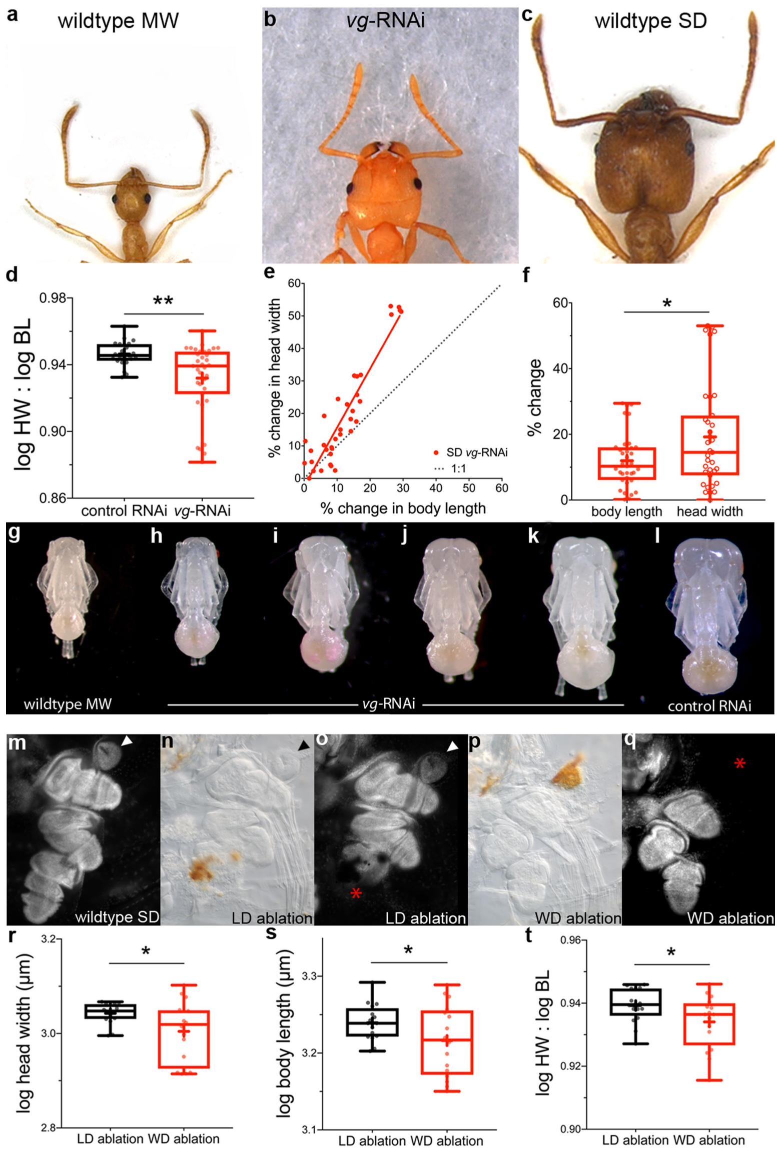
Extended Data Fig. 2 | Proliferation of rudimentary forewing discs in soldier-destined larvae. a–c, Immunohistochemistry of rudimentary forewing discs in *P. hyatti* soldier-destined larvae. DAPI (magenta) stains all nuclei within the cells of rudimentary forewing discs, and phospho-histone H3 (PH3; yellow) stains proliferating cells. The length of the larva from which the rudimentary forewing disc was dissected is given in the bottom right corner. All images are to scale. **c′** provides an increased magnification of **c**, to show nuclear co-localization of DAPI and PH3. Experiments were repeated twice.



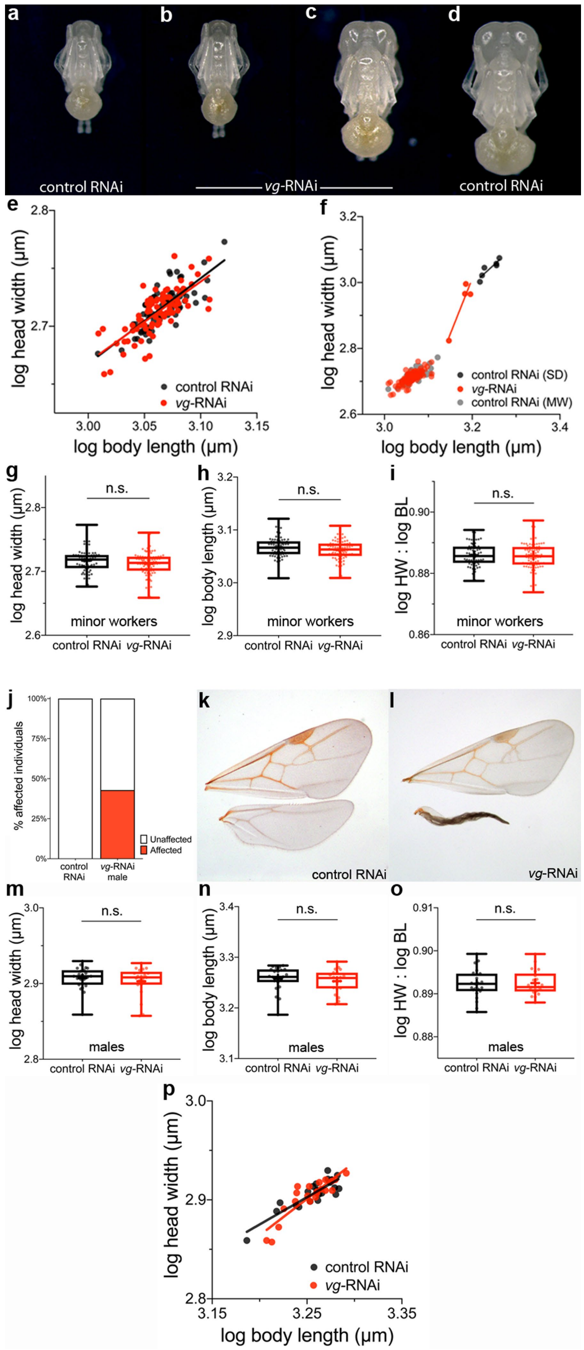
Extended Data Fig. 3 | Imaginal disc expression of *vg* is restricted to wing discs during larval development in *P. hyatti*. **a**–**c**, DAPI staining (**a**); *wingless* (*wg*) expression marking segment boundaries (**b**) and *vestigial* (*vg*) expression in the ventral nerve cord and in wing primordial cells in thoracic segments T2 and T3 (**c**) is shown for *P. hyatti* embryos. The three thoracic segments are labelled T1, T2 and T3. Experiments on embryos were repeated twice. **d**–**f**, Larval cartoons depicting leg discs and wing discs in queens (**d**), leg discs in minor workers (**e**) and leg discs and rudimentary wing discs in soldiers (**f**). Asterisks indicate absence of rudimentary wing discs. *vg* expression is indicated in purple. Lines at bottom right indicate the relative scale. **g**–**o**, *vg* expression is present in the larval (rudimentary) wing discs and is absent in the head and in leg discs of queens ($n = 17$) (**g**, **j**, **m**), minor workers ($n = 20$) (**h**, **k**, **n**) and soldiers ($n = 11$) (**i**, **l**, **o**). Black arrowheads indicate position of larval head and leg discs, and asterisks indicate absence of rudimentary wing discs. Experiments were repeated at least three times. **p**, Electrophoresis of PCR products obtained using *vg* and *EF1a* primers on wild-type soldier and minor-worker cDNA libraries, each constructed from three terminal-stage larvae. The 1 kb+ ladder is shown as reference and 1, 1/2, 1/4 and 1/8 represent serial dilutions of template cDNA. *vg* transcript (left) is detected in both soldiers and minor workers using semi-quantitative reverse-transcription PCR. The housekeeping *EF1a* transcript (right) is detected in both soldiers and minor workers, and levels are comparable between soldier and minor-worker cDNA libraries across dilutions. Negative controls (water with no template) show no contamination. For uncropped gel source data, see Supplementary Fig. 1. Experiments were repeated with independent biological replicates three times.



Extended Data Fig. 4 | *vg* RNAi reduces *vg* expression and induces apoptosis in rudimentary forewing discs. a, Electrophoresis of PCR products obtained using *vg* and *NADH* primers on soldier cDNA libraries each constructed from three terminal-stage larvae after *vg* RNAi or *yfp* RNAi (control RNAi) injection (Fig. 2a, red arrowhead). The 1 kb+ ladder is shown as reference and 1, 1/2, 1/4 and1/8 represent serial dilutions of template cDNA. *vg* transcript (left) is detected in both *vg* RNAi and control RNAi libraries using semi-quantitative reverse-transcription PCR. The housekeeping *NADH* transcript (right) is detected in both *vg* RNAi and control RNAi larvae, and levels are comparable between *vg* RNAi and control RNAi cDNA libraries across dilutions. Negative controls (water with no template) show no contamination. Experiments were repeated twice as independent biological replicates. For uncropped gel source data, see Supplementary Fig. 1. **b**, **c**, Apoptosis revealed by TUNEL assay in rudimentary forewing disc in control RNAi (**b**) and *vg* RNAi (**c**). Compared to control RNAi, *vg* RNAi induces apoptosis along the dorso-ventral margin of rudimentary forewing discs, where wild-type *vg* expression is normally strongest (**b**, **c**, black arrowhead). Experiments were repeated at least twice.

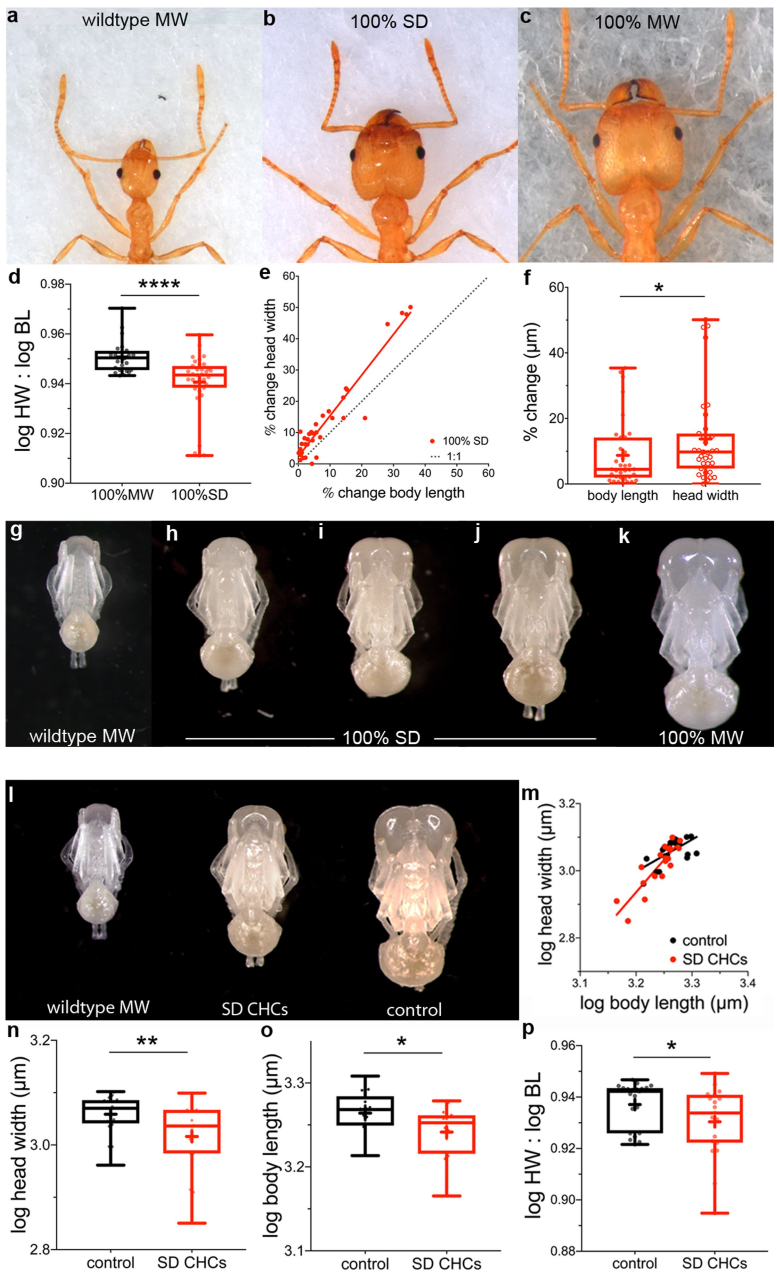


Extended Data Fig. 5 | Rudimentary forewing discs regulate size and disproportionate head-to-body scaling. **a**, Wild-type adult minor worker. **b**, *vg* RNAi intermediate adult. **c**, Wild-type soldier adult. Images to scale. **d**, Comparing ratios of log(head width (μm)) to log(body length (μm)) (log HW:log BL), between *yfp* RNAi (control RNAi, $n = 23$) and *vg* RNAi ($n = 35$); the box plot shows mean (+), interquartile range (bars), minimum-to-maximum values (whiskers); all points represent individual ants. Two-tailed Mann–Whitney *U*-test, $U = 219$, $^{**}P = 0.0031$. **e**, Percentage change in body length (μm) versus percentage change in head width (μm) of *vg* RNAi compared to a 1:1 line. Each point represents (absolute($HW - HW_{\text{control RNAi average}}$)/$HW_{\text{control RNAi average}}$) × 100 and/or (absolute($BL - BL_{\text{control RNAi average}}$)/$BL_{\text{control RNAi average}}$) × 100. **f**, Comparing the percentage change in body length (μm) and head width (μm) after *vg* RNAi ($n = 35$). The box plot shows mean (+), interquartile range (bars), maximum-to-minimum values (whiskers); all points represent individual ants. One-tailed Mann–Whitney *U*-test, $U = 470$, $^{*}P = 0.0477$. **g**, Wild-type minor worker. **l**, *yfp* RNAi (control RNAi) soldier. **g–l**, *vg* RNAi individuals showing a range of intermediates between minor worker and soldier (see Fig. 2i). Wild-type minor worker is shown for reference in **a**, **g**. All image comparisons are to scale. Experiments were repeated at least three times. **m–t**, Electrosurgical ablation of leg and rudimentary forewing discs in soldier-destined larvae (Fig. 2a, red arrowhead). **m**, Wild-type soldier, leg and rudimentary forewing discs. **n**, Site of leg disc cauterization shown by melanized cuticle. **o**, Ablation of leg disc (DAPI, red asterisk). **p**, Site of rudimentary forewing disc cauterization shown by melanized cuticle. **q**, Ablation of rudimentary forewing disc (DAPI, red asterisk). White or black arrowheads indicate the presence of rudimentary forewing discs. All images are to scale. **r**, Comparing log(head width (μm)) between leg disc ablation ($n = 16$) and rudimentary forewing disc ablation ($n = 16$); one-tailed Mann–Whitney *U*-test, $U = 82$, $^{*}P = 0.0432$. **s**, Comparing log(body length (μm)) between leg disc ablation ($n = 16$) and rudimentary forewing disc ablation ($n = 16$). One-tailed unequal variance *t*-test, $t = 1.77$, d.f. = 22.34, $^{*}P = 0.045$. **t**, Comparing ratio of log(head width (μm)) to log(body length(μm)), between leg disc ablation ($n = 16$) and rudimentary forewing disc ablation ($n = 16$). One-tailed unpaired *t*-test, $t = 2.07$, d.f. = 30, $^{*}P = 0.0234$. The box plots in **d**, **f**, **r–t** show mean (+), interquartile range (bars) and minimum-to-maximum values (whiskers); all points represent individual ants. Experiments were repeated at least twice.

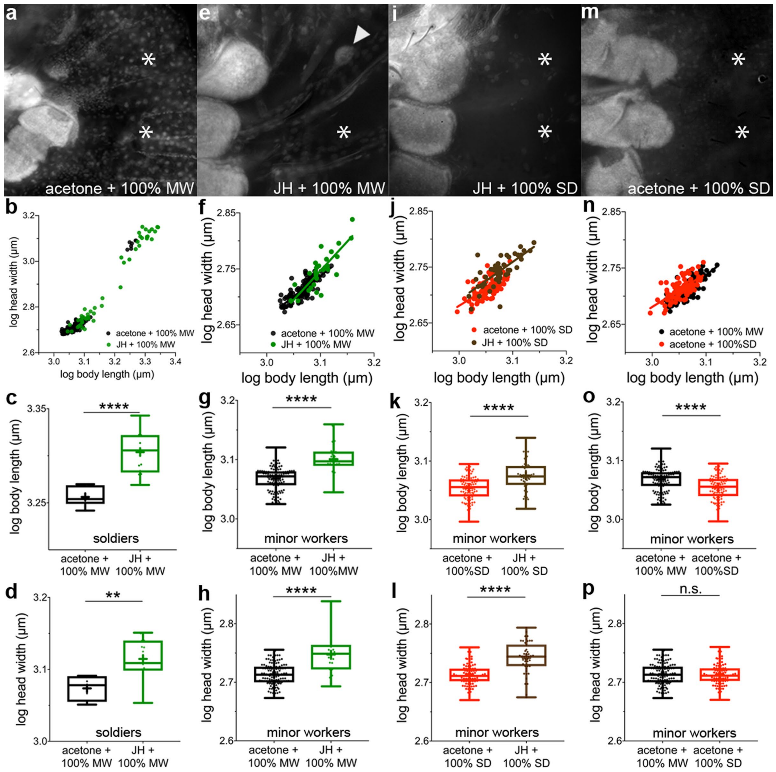


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | The function of rudimentary forewing discs in regulating disproportionate head-to-body scaling is specific to the soldier subcaste. *vg* RNAi on bipotential larvae (Fig. 2a, orange arrowhead). **a**, *yfp* RNAi (control RNAi) minor worker. **b**, *vg* RNAi minor worker. **c**, *vg* RNAi intermediate. **d**, *yfp* RNAi (control RNAi) soldier. Image comparisons to scale. **e**, log(body length (μm)) versus log(head width (μm)), comparing minor worker head-to-body scaling of *yfp* RNAi (control RNAi, $n=69$) and *vg* RNAi ($n=84$). ANCOVA: slope, $F=0.162$, d.f. = 149, $P=0.69$; *y* intercept, $F=0.4755$, d.f. = 150, $P=0.49$). **f**, log(body length (μm)) versus log(head width (μm)), comparing head-to-body scaling of *yfp* RNAi (control RNAi; $n=6$) and *vg* RNAi ($n=4$) ants that initiated soldier development. ANCOVA: $F=7.44$, d.f. = 6, $P=0.0343$. **g**, Comparing log(head width (μm)) between control RNAi ($n=69$) and *vg* RNAi ($n=84$) minor workers. Two-tailed unpaired *t*-test: $t=1.62$, d.f. = 151, $P=0.1074$. **h**, Comparing log(body length (μm)) between control RNAi ($n=69$) and *vg* RNAi ($n=84$) minor workers. Two-tailed unpaired *t*-test: $t=1.54$, d.f. = 151, $P=0.1249$. **i**, Comparing ratios of log(head width (μm)) to log(body length (μm)) between control RNAi ($n=69$) and *vg* RNAi ($n=84$) minor workers. Two-tailed unpaired *t*-test: $t=0.26$, d.f. = 151, $P=0.80$. Experiments were repeated at least three times. **j**–**p**, *vg* RNAi on male-destined larvae (Fig. 2a, green arrowhead). **j**, Bar graph showing percentage of individual ants affected (red) after *yfp* RNAi (control RNAi; $n=0$ out of 23) and *vg* RNAi ($n=9$ out of 21). Two-tailed Fisher's exact test: $P=0.0004$. **k**, **l**, Wings from *yfp* RNAi (control RNAi) (**k**) and *vg* RNAi (**l**) males. Image comparisons are to scale. **m**–**o**, Head-to-body scaling of *yfp* RNAi (control RNAi; $n=23$) and *vg* RNAi males ($n=21$), comparing log(head width (μm)), two-tailed Mann–Whitney *U*-test, $U=222.5$, $P=0.6626$ (**m**); log(body length (μm)), two-tailed Mann–Whitney *U*-test, $U=198$, $P=0.3157$ (**n**); and ratio of log(head width (μm)) to log(body length (μm)) (**o**), between control RNAi males and *vg* RNAi males, two-tailed unpaired *t*-test, $t=0.18$, d.f. = 42, $P=0.86$. **p**, log(body length (μm)) versus log(head width (μm)) comparison of head-to-body scaling, between *yfp* RNAi (control RNAi; $n=23$) and *vg* RNAi males ($n=21$). ANCOVA: slope, $F=3.111$, d.f. = 40, $P=0.0854$; *y* intercept, $F=0.076$, d.f. = 41, $P=0.7837$. Experiments were repeated at least three times. The box plots in **g**–**i** and **m**–**o** show mean (+), interquartile range (bars) and minimum-to-maximum values (whiskers); all points represent individual ants.



Extended Data Fig. 7 | Soldier inhibitory pheromone regulates size and disproportionate head-to-body scaling. a–k, Effect of social inhibition on soldier-destined larvae (Fig. 2a, red arrowhead). **a**, Wild-type adult minor worker. **b**, Intermediate adult resulting from soldier-destined larvae being raised in colonies composed of 100% soldiers (high inhibition). **c**, Representative adult soldier resulting from soldier-destined larvae being raised in colonies composed of 100% minor workers (no inhibition). **d**, Comparing ratios of log(head width (μm)) to log(body length (μm)) between 100% minor worker ($n = 24$) and 100% soldier ($n = 35$). The box plot shows mean (+), interquartile range (bars) and minimum-to-maximum values (whiskers); all points represent individual ants. Two-tailed Mann–Whitney *U*-test, $U = 155$, $^{****}P < 0.0001$. **e**, Percentage change in body length (μm) versus percentage change in head width (μm) of 100% soldiers, compared to a 1:1 line. Each point represents (absolute($HW - HW_{\text{100\% minor worker average}}$)/$HW_{\text{100\% minor worker average}}$) $\times$ 100 and/or (absolute($BL - BL_{\text{100\% minor worker average}}$)/$BL_{\text{100\% minor worker average}}$) $\times$ 100. **f**, Comparing percentage change in body length (μm) and head width (μm) following high inhibition (100% soldiers; $n = 35$). One-tailed Mann–Whitney *U*-test, $U = 421$, $^{*}P = 0.0121$. **g**, Wild-type minor worker. **h–j**, 100% soldier-raised ants showing a range of intermediates between minor workers and soldiers (see Fig. 3e). **k**, 100% minor-worker control soldier. All image comparisons are to scale. Experiments were repeated at least three times. **l–p**, Application of soldier cuticular hydrocarbon extract (CHCs) to soldier-destined larvae (Fig. 2a, red arrowhead). **l**, Wild-type minor worker and individual ants treated with soldier CHCs and hexane solvent (control). **m**, Comparing slopes of hexane solvent control ($n = 21$) and soldier CHCs ($n = 19$). ANCOVA, $F = 6.84$, d.f. $= 36$, $P = 0.0129$. **n**, log(head width (μm)); one-tailed Mann–Whitney *U*-test, $U = 114$, $^{**}P = 0.0099$. **o**, log(body length (μm)); one-tailed Mann–Whitney *U*-test, $U = 117$, $^{*}P = 0.0126$. **p**, Ratio of log(head width (μm)) to log(body length (μm)); one-tailed Mann–Whitney *U*-test, $U = 125$, $^{*}P = 0.0221$. Wild-type minor worker is shown for reference in **a**, **g**, **l**. The box plots in **d**, **f**, **n–p** show mean (+), interquartile range (bars), minimum-to-maximum values (whiskers); all points represent individual ants.



Extended Data Fig. 8 | Juvenile hormone and inhibitory pheromone regulate disc-dependent disproportionate head-to-body scaling and disc-independent proportional head:body scaling. Effect of juvenile-hormone activation and social inhibition on bipotential larvae (Fig. 2a, orange arrowhead). **a**, Absence of rudimentary wing discs in minor worker larvae exposed to solvent control with no inhibition ('acetone + 100% MW'). **a**, **e**, **i**, **m**, Arrowheads indicate the presence of rudimentary wing discs and asterisks indicate the absence of rudimentary wing discs. **b**, Plot of head-to-body scaling of acetone + 100% minor worker, and juvenile-hormone activation with no inhibition ('JH + 100% MW'); the majority of larvae treated with juvenile hormone develop into soldiers, and some develop into large minor workers. **c**, **d**, Comparing between acetone + 100% minor worker ($n = 7$) and juvenile hormone + 100% minor worker ($n = 17$) treatments of individuals that developed into the soldier size distribution. **c**, log(body length (µm)); two-tailed unpaired t-test; $t = 5.25$, d.f. = 22, $^{****}P < 0.0001$. **d**, log(head width (µm)); two-tailed unpaired t-test, $t = 3.50$, d.f. = 22, $^{**}P = 0.002$. **e**, Initiation of growth of rudimentary forewing discs in minor worker larvae exposed to juvenile hormone + 100% minor worker. **f**, Plot of head-to-body scaling between acetone + 100% minor worker ($n = 114$) and juvenile hormone + 100% minor worker ($n = 29$) of individuals that developed into the minor-worker size distribution. ANCOVA: $F = 7.12$, d.f. = 139, $P = 0.0085$. **g**, **h**, Comparing between acetone + 100% minor worker ($n = 114$) and juvenile hormone + 100% minor worker ($n = 29$) treatments of individuals that developed into the minor-worker size distribution. **g**, log(body length (µm)); two-tailed Mann–Whitney U-test, $U = 463$, $^{****}P < 0.0001$. **h**, log(head width (µm)); two-tailed unequal variance t-test, $t = 5.19$, d.f. = 32, $^{****}P < 0.0001$. **i**, Absence of growth of rudimentary wing discs in minor worker larvae exposed to juvenile-hormone activation with high inhibition ('JH + 100% SD'). **j**, Plot of head-to-body scaling between solvent control with high inhibition ('acetone + 100% SD'; $n = 88$) and juvenile hormone + 100% soldiers ($n = 46$) of individuals that developed into the minor-worker size distribution. ANCOVA: slope, $F = 0.54$, d.f. = 130, $P = 0.47$; y intercept, $F = 27.2$, d.f. = 131, $P < 0.0001$. **k**, **l**, Comparing between acetone + 100% soldiers ($n = 88$) and juvenile hormone + 100% soldiers ($n = 46$) treatments of individuals that developed into the minor-worker size distribution. **k**, log(body length (µm)); two-tailed unequal variance t-test, $t = 4.43$, d.f. = 69, $^{****}P < 0.0001$. **l**, log(head width (µm)); two-tailed unequal variance t-test, $t = 6.69$, d.f. = 68, $^{****}P < 0.0001$. **m**, Absence of rudimentary wing disc growth in minor worker larvae exposed to acetone + 100% soldiers. **n**, Plot of head-to-body scaling between acetone + 100% minor workers ($n = 114$) and acetone + 100% soldiers ($n = 88$) treatments of individuals that developed into the minor-worker size distribution. ANCOVA: slope, $F = 2.84$, d.f. = 198, $P = 0.0937$; y intercept, $F = 20.77$, d.f. = 199, $P < 0.0001$. **o**, **p**, Comparing between acetone + 100% minor workers ($n = 114$) and acetone + 100% soldiers ($n = 88$) treatments of individuals that developed into the minor-worker size distribution. **o**, log(body length (µm)); two-tailed Mann–Whitney U-test, $U = 3178$, $^{****}P < 0.0001$. **p**, log(head width (µm)); two-tailed unpaired t-test, $t = 0.28$, d.f. = 200, $P = 0.7823$. Bonferroni correction was applied to comparison of log(head width) and log(body length). The box plots in **c**, **d**, **g**, **h**, **k**, **l**, **o**, **p** show mean (+), interquartile range (bars) and minimum-to-maximum values (whiskers); all points represent individual ants. Plots in **b**, **f**, **j**, **n** show linear regressions in which the x axis is log(body length (µm)) and the y axis is log(head width (µm)). All images are to scale. Experiments were repeated at least three times.

Extended Data Table 1 | Description and references for presence or absence of soldier subcaste and discrete variation in rudimentary wing disc size across 21 ant species

Species	Presence of a soldier (major worker) subcaste	Reference	Inter-subcaste variation in rudimentary wing disc size?	Description of variation in rudimentary wing disc size	Reference
Pheidole rhea	Yes	Wilson[8], Rajakumar et al.[10], Huang & Wheeler[57]	Yes	MW= no visible discs SD = 2 pairs of discs XSD = 2 pairs of large discs	Rajakumar et al.[10]
Pheidole megacephala	Yes	Wilson[8]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Rajakumar et al.[10], Sameshima et al.[11]
Pheidole spadonia	Yes	Wilson[8], Huang & Wheeler[57]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Rajakumar et al.[10]
Pheidole pilifera	Yes	Wilson[8]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Rajakumar et al.[10]
Pheidole tysoni	Yes	Wilson[8]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Rajakumar et al.[10]
Pheidole bicarinata	Yes	Wilson[8], Wheeler & Nijhout[9]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Wheeler & Nijhout[9]
Pheidole moerens	Yes	Wilson[8]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Rajakumar et al.[10]
Pheidole obtusospinosa	Yes	Wilson[8], Rajakumar et al.[10], Huang & Wheeler[57]	Yes	MW= no visible discs SD = 1 pairs of forewing discs XSD = 2 pairs of large discs	Rajakumar et al.[10], Present study
Pheidole morrisi	Yes	Wilson[8]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Abouheif & Wray[7], Rajakumar et al.[10], Shbailat & Abouheif[58]
Pheidole hyatti	Yes	Wilson[8]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Rajakumar et al.[10], Present study
Pheidole vallicola	Yes	Wilson[8]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Rajakumar et al.[10]
Pheidole dentata	Yes	Wilson[8]	Yes	MW= no visible discs SD = 1 pair of forewing discs	Rajakumar et al.[10]
Tetramorium caespitum	No	Bharti & Kumar[59], Bolton[60]	No	W = 2 pairs of small wing discs	Shbailat & Abouheif[58]
Crematogaster cerasi	No	Morgan & Mackay[61]	No	W = 2 pairs of small wing pads	Abouheif & Wray[7], Shbailat & Abouheif[58]
Monomorium trageri	No	Dubois[62]	No	W = no visible discs	Present study
Solenopsis geminata	Yes	Tschinkel[30], Wilson[63]	Yes	MW = 2 pairs of small wing pads SD = 2 pairs of large wing discs	Present study
Myrmica americana	No	Weber[64]	No	W = 2 pairs of small wing discs	Abouheif & Wray[7]
Lasius niger	No	Wilson[65]	No	W = 2 pairs of small wing discs	Shbailat & Abouheif[58]
Formica pallidefulva (described as *Neoformica nitidiventris* in[7])	No	Trager, MacGown & Trager[66]	No	W = 2 pairs of small wing discs	Abouheif & Wray[7]
Camponotus floridanus	Yes	Alvarado[29], Wilson[3]	Yes	MW = 2 pairs of small wing pads SD = 2 pairs of large wing discs	Present study
Mysterium obertheuri	No	Béhague et al.[67]	No	W = 2 pairs of wing discs	Béhague et al.[67]

MW, minor worker; SD, soldier; W, worker; XSD, supersoldier. Refer to Extended Data Fig. 1a for the phylogeny. Data were obtained from previous publications[7–11,57–67].

https://doi.org/10.1038/s41586-018-0543-y

Clearance of senescent glial cells prevents tau-dependent pathology and cognitive decline

Tyler J. Bussian[1,3], Asef Aziz[2,3], Charlton F. Meyer[2], Barbara L. Swenson[2], Jan M. van Deursen[1,2] & Darren J. Baker[1,2]*

Cellular senescence, which is characterized by an irreversible cell-cycle arrest[1] accompanied by a distinctive secretory phenotype[2], can be induced through various intracellular and extracellular factors. Senescent cells that express the cell cycle inhibitory protein p16^{INK4A} have been found to actively drive naturally occurring age-related tissue deterioration[3,4] and contribute to several diseases associated with ageing, including atherosclerosis[5] and osteoarthritis[6]. Various markers of senescence have been observed in patients with neurodegenerative diseases[7–9]; however, a role for senescent cells in the aetiology of these pathologies is unknown. Here we show a causal link between the accumulation of senescent cells and cognition-associated neuronal loss. We found that the *MAPT^{P301S}PS19* mouse model of tau-dependent neurodegenerative disease[10] accumulates p16^{INK4A}-positive senescent astrocytes and microglia. Clearance of these cells as they arise using *INK-ATTAC* transgenic mice prevents gliosis, hyperphosphorylation of both soluble and insoluble tau leading to neurofibrillary tangle deposition, and degeneration of cortical and hippocampal neurons, thus preserving cognitive function. Pharmacological intervention with a first-generation senolytic modulates tau aggregation. Collectively, these results show that senescent cells have a role in the initiation and progression of tau-mediated disease, and suggest that targeting senescent cells may provide a therapeutic avenue for the treatment of these pathologies.

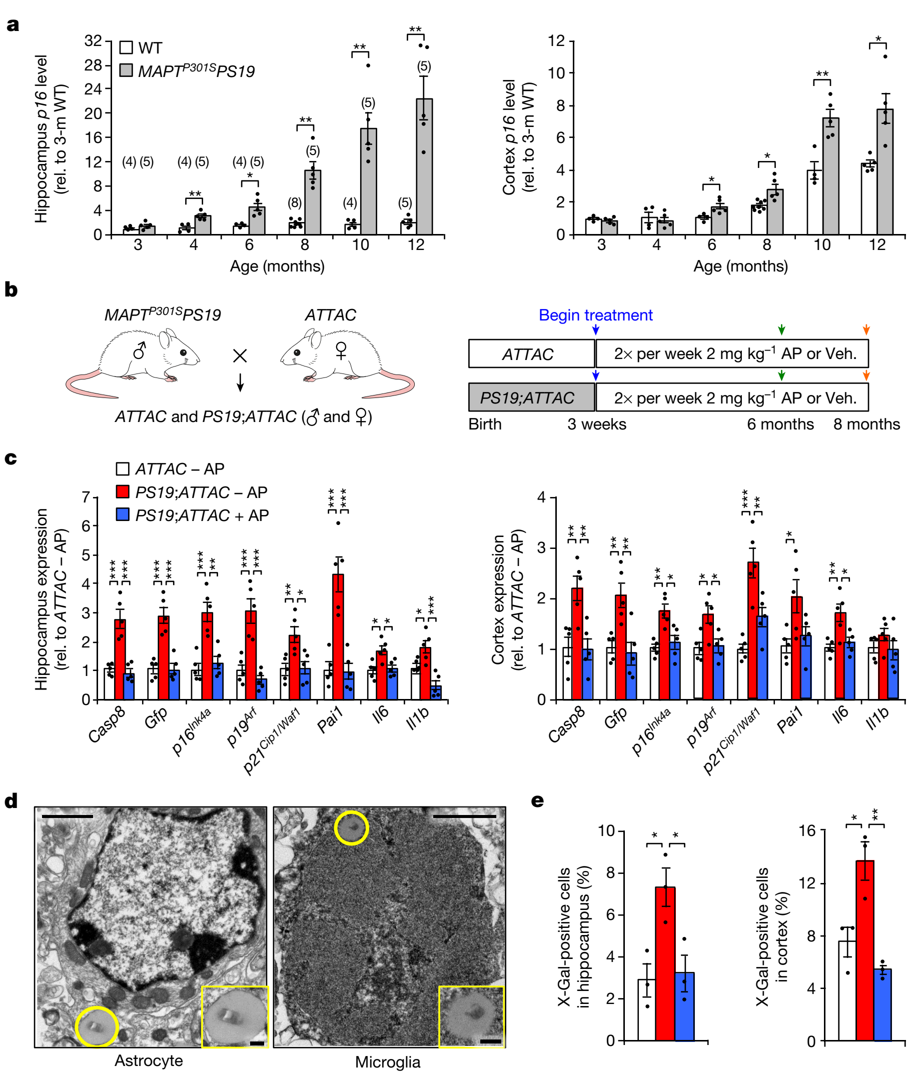


Fig. 1 | Senescent astrocytes and microglia that accumulate in the brains of *MAPT^{P301S}PS19* mice can be removed using the *INK-ATTAC* transgene. **a**, RT–qPCR analysis of *p16^{Ink4a}* expression in the hippocampus (left) and cortex (right) of wild-type (WT) and *MAPT^{P301S}PS19* mice. The number of mice for each column are indicated in parentheses in the left-hand graph, 2 independent experiments; normalized to the three-month wild-type group. **b**, Study design for the clearance of senescent cells in *PS19;ATTAC* mice. Veh., vehicle. **c**, RT–qPCR analysis of the expression of senescence markers in the hippocampus (left) and cortex (cortex) of six-month-old male mice, treated with either vehicle (−AP) or AP20187 (+AP). $n = 5$ mice per group; normalized to the *ATTAC* −AP group. **d**, Electron micrograph showing an X-Gal-positive astrocyte (left) and microglia (right) after SA-β-Gal staining, from a six-month-old vehicle-treated *PS19;ATTAC* male mouse. **e**, Quantification of cells containing X-Gal crystals in the hippocampus (left) or the cortex (right) of six-month-old male mice. $n = 3$ male mice per group, 2 independent experiments; the bars are coloured as in **c**. Scale bars, 1 μm (**d**) and 200 nm (**d**, insets). Data are mean ± s.e.m. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$ (unpaired two-sided *t*-tests with Welch's correction (**a**) and one-way ANOVA with Tukey's multiple comparisons test (**c**, **e**)). Exact *P* values can be found in the accompanying Source Data.

[1]Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, MN, USA. [2]Department of Pediatric and Adolescent Medicine, Mayo Clinic, Rochester, MN, USA. [3]These authors contributed equally: Tyler J. Bussian, Asef Aziz. *e-mail: baker.darren@mayo.edu

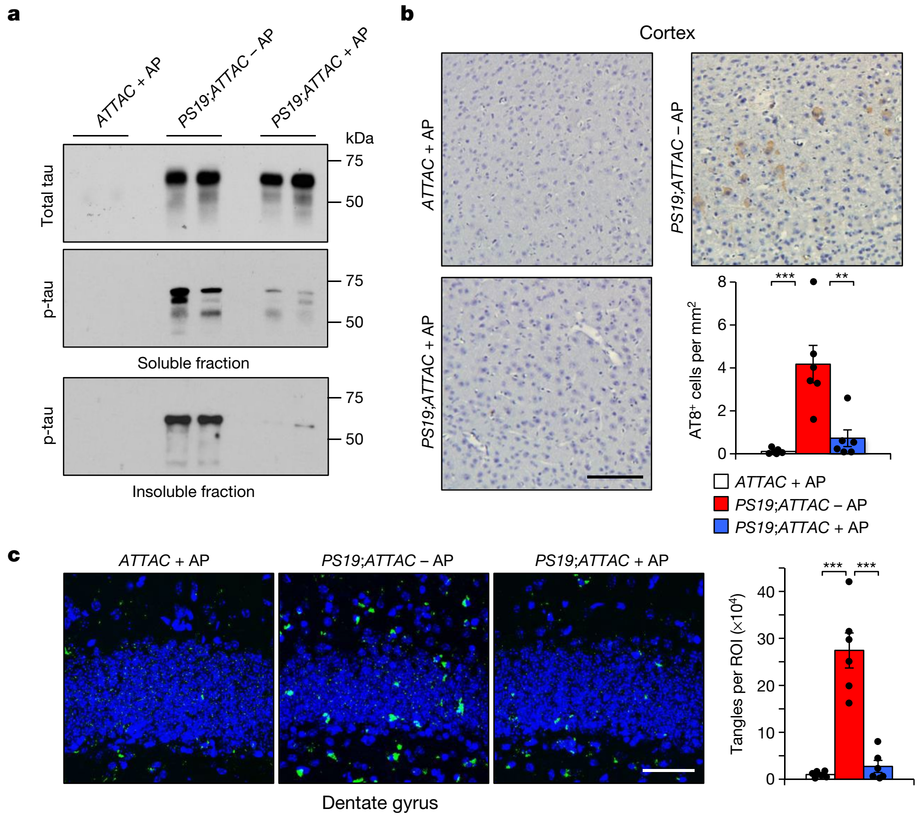


Fig. 2 | Senescent cells promote insoluble tau aggregates. a, Representative western-blot analysis of the whole brain of six-month-old mice for soluble tau (top), soluble phosphorylated tau (S202/T205; middle) and insoluble phosphorylated tau (S202/T205; bottom). ≥3 independent experiments. **b**, Immunostaining and quantification of cortex sections from six-month-old mice for phosphorylated tau (S202/T205) protein aggregates. $n = 6$ mice per group, 3 independent experiments. **c**, Thioflavin-S staining and quantification of neurofibrillary tangles located within the dentate gyrus of eight-month-old mice. $n = 6$ mice per group, 2 independent experiments; normalized to the *ATTAC* + AP group. The bar graph is coloured as in **b**. ROI, region of interest. Scale bars, 100 μm (**b**) and 50 μm (**c**). Data are mean ± s.e.m. $^{**}P < 0.01$; $^{***}P < 0.001$ (one-way ANOVA with Tukey's multiple comparisons test). Exact P values can be found in the accompanying Source Data. For gel source data, see Supplementary Fig. 1.

Senescent cells accumulate with ageing and have been shown to contribute to tissue dysfunction[11], although their role in neurodegenerative disease is still unknown. To address this key open question, we selected the transgenic mouse line *MAPT*^{P301S}*PS19* (hereafter *PS19*), which—under the regulation of the mouse prion promoter—expresses high levels of mutant human tau specifically in neurons[10]. The model is characterized by gliosis, neurofibrillary tangle (NFT) deposition, neurodegeneration and loss of cognitive function. Pathology typically initiates in the hippocampus and radiates outwards to the neocortex[10]. First, we performed quantitative PCR with reverse transcription (RT–qPCR) for *p16*^{Ink4a} (also known as *Cdkn2a*) on isolated hippocampi and cortices from wild-type and *PS19* littermates. A significant increase in *p16*^{Ink4a} expression was seen in the *PS19* mice beginning at four months of age in the hippocampus and at six months of age in the cortex (Fig. 1a), which precedes the onset of NFT deposition[10]. Importantly, increased expression of *p16*^{Ink4a} correlated with the expression of widely established senescence markers (Extended Data Fig. 1), indicating that senescent cells accumulate at sites of pathology in the *PS19* model.

To investigate the role of senescent cells in the development of disease, we crossed the *INK-ATTAC* transgene (hereafter *ATTAC*) to the *PS19* strain to eliminate *p16*^{Ink4a}-expressing senescent cells through the twice-weekly administration of AP20187 (hereafter AP)[3,4] from weaning age (Fig. 1b). Hippocampi and cortices isolated from six-month-old vehicle-administered *PS19;ATTAC* mice displayed an increased level of the *ATTAC* transgene, as measured by the expression of *Casp8* and *Gfp* (Fig. 1c, Extended Data Fig. 2). Levels of senescence indicators, including the cell-cycle regulators *p16*^{Ink4a}, *p19*^{Arf} (also known as *Cdkn2a*) and *p21*$^{Cip1/Waf1}$ (also known as *Cdkn1a*) and the pro-inflammatory genes *Pai1* (also known as *Serpine1*), *Il6* and *Il1b*, were also increased (Fig. 1c, Extended Data Fig. 2). Administration of AP to *PS19;ATTAC* mice maintained the expression of these genes at a level comparable to that of control mice (Fig. 1c, Extended Data Fig. 2). Importantly, treatment of *ATTAC* mice that lacked the *PS19* transgene with AP had no effect on the expression of these markers (Extended Data Fig. 2). Thus, AP administration effectively and selectively cleared senescent cells in the hippocampus and cortex of *PS19;ATTAC* mice.

To understand the mechanistic contribution of senescence to tau-mediated pathology, we sought to identify the specific cell types that were becoming senescent. First, we stained cortices and hippocampi from six-month-old vehicle-treated *ATTAC* and *PS19;ATTAC* mice and AP-treated *PS19;ATTAC* mice for senescence-associated-β-galactosidase (SA-β-Gal)[12] and screened for cells that contained 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal) crystals by transmission electron microscopy[13]. We found that cells that clearly and morphologically resembled astrocytes or microglia contained X-Gal crystals, irrespective of the treatment group from which they arose (Fig. 1d). By contrast, no crystals were found in any clearly identifiable neurons (Extended Data Fig. 3). Compared with control mice, vehicle-treated *PS19;ATTAC* mice had nearly double the number of cells containing X-Gal crystals in both the hippocampus and the cortex (Fig. 1e), whereas AP-treated *PS19;ATTAC* mice had a similar incidence of X-Gal crystals as control mice (Fig. 1e). To validate that senescence was affecting astrocytes and microglia, we performed fluorescence-activated cell sorting (FACS) on six-month-old wild-type and *PS19* mice (Extended Data Fig. 4a). Isolated astrocytes and microglia had increased expression of senescence-associated genes, including *p16*^{Ink4a} (Extended Data Fig. 4b, c). A similar induction was not observed in oligodendrocytes or neuron-enriched CD56^{+} cells (Extended Data Figs. 4d, e, 5), supporting the conclusion that senescence occurs primarily in astrocytes and microglia of *PS19* mice.

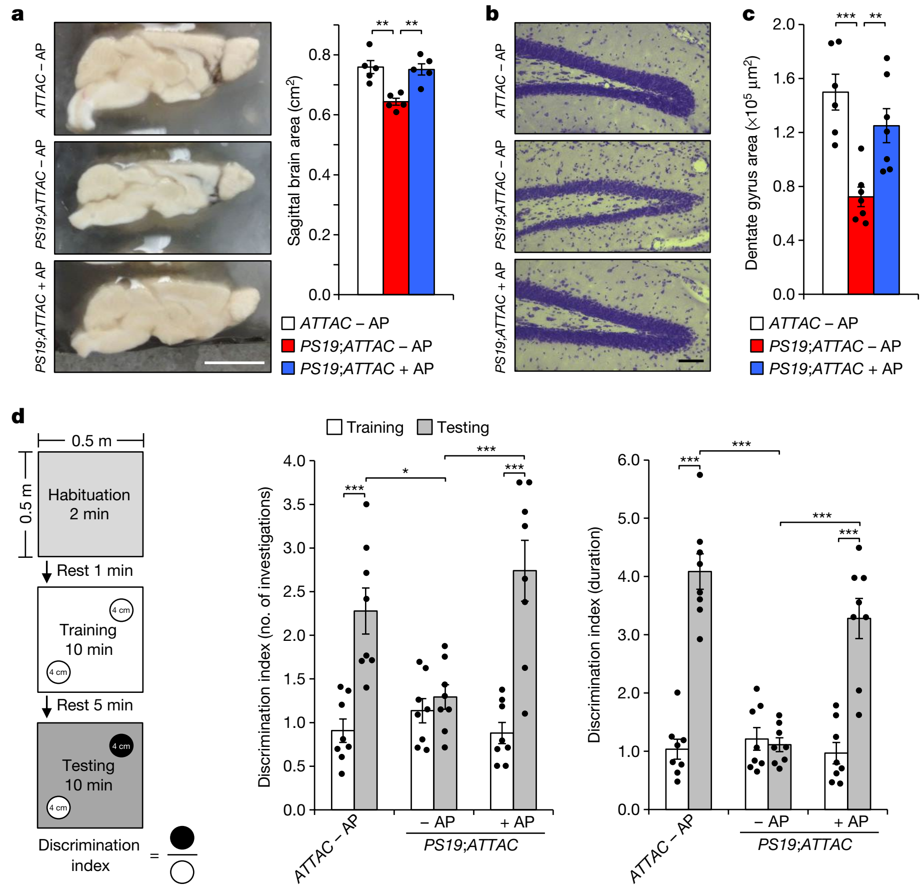


Fig. 3 | Senescent cells drive neurodegenerative disease. **a**, Sagittal midline brain area images (left) and quantification (right) of eight-month-old mice. $n = 5$ males per group, 2 independent experiments. **b**, Nissl stains of the dentate gyrus (left) and quantification (right) from eight-month-old mice. **c**, Average area of the dentate gyrus (measuring the pyramidal neuron layer) from serial, coronal NeuN-stained free-floating sections. $n = 6$ *ATTAC* –AP and $n = 7$ *PS19*;*ATTAC* – AP and *PS19*;*ATTAC* + AP mice, 2 independent experiments. **d**, Setup of the novel-object recognition experiment (left) and average ratio for the number of investigations (middle) and duration of those investigations (right). $n = 8$ female mice per group. Scale bars, 0.5 cm (**a**) and 100 μm (**b**). Data are mean ± s.e.m. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$ (one-way ANOVA with Tukey's multiple comparisons test (**a**, **c**) and two-way ANOVA with Tukey's multiple comparisons test (**d**)). Exact *P* values can be found in the accompanying Source Data.

To verify that the administration of AP selectively targeted senescent cells, we prepared in vitro cultures of primary microglia and astrocytes isolated from *ATTAC* mice. These cultures were not sensitive to AP-mediated elimination in the absence of senescence-inducing stimuli (Extended Data Fig. 6). Furthermore, for *ATTAC* transgenic mice in vivo, the short-term administration of AP did not promote excessive cellular death (Extended Data Fig. 7a), and extended treatment with AP did not result in increased proliferation of microglia (Extended Data Fig. 7b).

PS19 mice present progressive gliosis with disease progression[10]. To assess whether the administration of AP affected this process, RT–qPCR was performed on the hippocampi of six-month-old mice for markers of astrocytes (*Gfap* and *S100b*) and microglia (*Cd11b*, also known as *Itgam*). Vehicle-treated *PS19*;*ATTAC* mice had an approximately two- to threefold induction in these markers, whereas AP-treated *PS19*;*ATTAC* mice expressed these markers at a similar level to control mice (Extended Data Fig. 8a, b). Immunohistochemistry staining for GFAP and IBA1 confirmed these observations (Extended Data Fig. 8c, d). Taken together, these results suggest that both senescent glial cells and gliosis are eliminated upon the administration of AP in *PS19*;*ATTAC* mice.

A distinguishing characteristic of *PS19* mice is the development of aggregates consisting of hyperphosphorylated tau protein by six months of age[10]. To assess whether tau aggregation was affected by senescence clearance, we measured the levels of soluble total and phosphorylated tau (S202/T205) in addition to the level of insoluble phosphorylated tau in vehicle-treated *PS19*;*ATTAC* and AP-treated *ATTAC* and *PS19*;*ATTAC* mice. As expected[10], vehicle-treated *PS19*;*ATTAC* mice displayed increased levels both of soluble total and phosphorylated tau and of insoluble phosphorylated tau (Fig. 2a, Extended Data Fig. 9a, b). AP-treated *PS19*;*ATTAC* mice had identical levels of soluble total tau to vehicle-treated *PS19*;*ATTAC* mice (Fig. 2a), indicating that overexpression of tau from the transgene was maintained. Treatment of *PS19*;*ATTAC* mice with AP significantly reduced the amount of phosphorylated tau in both the soluble and the insoluble fractions (Fig. 2a, Extended Data Fig. 9b). Immunohistochemistry staining for phospho-tau modifications at S202/T205, T231 and S396 confirmed that the clearance of senescent cells attenuated tau phosphorylation at several residues that are relevant for tau aggregation (Fig. 2b, Extended Data Fig. 9c). Additionally, the staining of brain sections of eight-month-old mice from these same groups with thioflavin-S revealed that NFT deposition in the dentate gyrus—the site of neurogenesis in the hippocampus that is traditionally associated with memory formation and cognition[14]—was substantially reduced when senescent cells were removed (Fig. 2c). Collectively, these results indicate that the accumulation of senescent cells promotes the formation of hyperphosphorylated tau aggregates.

PS19 mice show neurodegeneration by eight months of age[10]. As NFT deposition was attenuated upon treatment with AP in both the cortex and the hippocampus of *PS19*;*ATTAC* mice, we performed assessments for degeneration in these areas. The overt brain size of vehicle-treated *PS19*;*ATTAC* mice was reduced compared to both

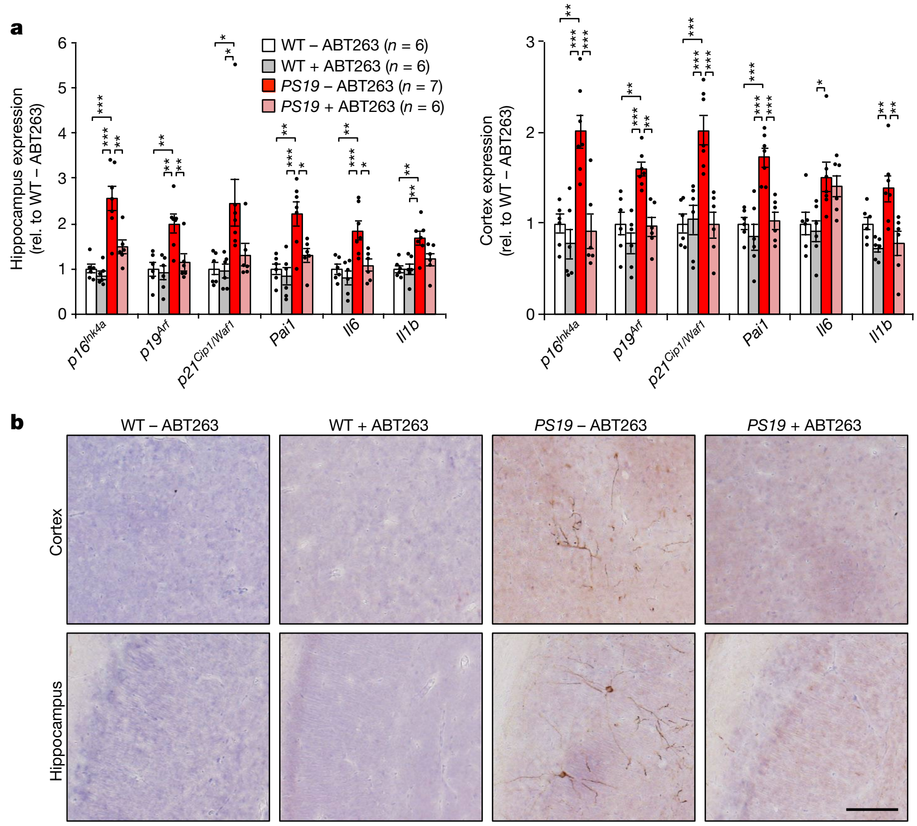


Fig. 4 | ABT263 can modulate senescent cells and attenuate tau phosphorylation. **a**, RT–qPCR analysis of the expression of senescence markers from the hippocampus (left) and the cortex (right) of six-month-old mice treated with either vehicle (−ABT263) or ABT263 (+ABT263). Number of mice is as indicated; normalized to the WT − ABT263 group. **b**, Representative immunostaining of cortex (top) and hippocampal (bottom) sections for phosphorylated tau (S202/T205) protein aggregates. $n = 4$ mice per group, 2 independent experiments. Scale bar, 100 μm. Data are mean ± s.e.m. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$ (one-way ANOVA with Tukey's multiple comparisons test). Exact P values can be found in the accompanying Source Data.

ATTAC and AP-treated *PS19*;*ATTAC* mice (Fig. 3a). In addition, we observed localized neurodegeneration in the dentate gyrus of the hippocampus through Nissl staining in vehicle-treated *PS19*;*ATTAC* mice (Fig. 3b). The administration of AP prevented thinning of the dentate gyrus and increased neuron density. Sequential coronal sectioning and NeuN staining revealed that the dentate gyrus was significantly reduced in area in vehicle-treated *PS19*;*ATTAC* mice (Fig. 3c), further demonstrating that senescent cells promote neurodegeneration in *PS19* mice.

To test whether the effects observed upon administration of AP resulted in improved cognitive function, we performed novel-scent discrimination assessments to test for changes in short-term memory[15] (for experimental setup, see Fig. 3d). Whereas AP-treated *ATTAC* mice were more inquisitive towards the novel scent during the testing phase, vehicle-treated *PS19*;*ATTAC* mice were not (Fig. 3d). By contrast, AP-treated *PS19*;*ATTAC* mice behaved nearly identically to control (AP-treated *ATTAC*) mice, indicating that the elimination of senescent cells mitigated the short-term memory loss observed in vehicle-treated *PS19*;*ATTAC* mice. Notably, the overall distance travelled by mice in all groups was unchanged (data not shown), and similar results were obtained in novel-object discrimination tests that had the same setup but used visual cues instead of scents (Extended Data Fig. 10). Thus, these results demonstrate that senescent cells drive neurodegeneration and loss of cognition in *PS19* mice.

Last, we tested whether the pharmacological elimination of senescent cells with the senolytic ABT263 (navitoclax)[5,6,16,17] resulted in similar effects to our genetic interventions in *PS19* mice. Recent work has demonstrated a therapeutic effect in orthotopically implanted glioblastomas of the peripheral administration of ABT263[18]. Wild-type and *PS19* mice were treated with a repeating schedule of ABT263, beginning at weaning age and continuing until the mice reached six months of age. Notably, this treatment prevented the upregulation of senescence-associated genes (Fig. 4a) and attenuated tau phosphorylation in *PS19* mice (Fig. 4b), indicating that senolytic interventions can recapitulate key observations from transgenic mouse models of senescent-cell ablation.

The mechanistic contribution of cells with features reminiscent of senescence to the pathophysiology of neurodegenerative diseases has been a common question in recent years[7–9,19–21]. Furthermore, recent work has suggested that senescent cells may contribute to the pathology of Parkinson's disease in both mice and humans[22]. Here we show that the continuous clearance of *p16*^{Ink4a}-expressing senescent cells before disease onset in a model of aggressive tauopathy has a marked effect on various aspects of disease progression, including gliosis, NFT formation, neurodegeneration and cognitive decline. Senescent-cell clearance also has a notable effect on the accumulation of phosphorylated tau protein in both the soluble and insoluble fractions. The amount of total soluble tau was unchanged in AP-treated *PS19*;*ATTAC* mice (Fig. 2a), indicating that the aberrant hyperphosphorylation of tau protein and subsequent aggregation into NFTs is mediated by extracellular signalling from *p16*^{Ink4a}-expressing senescent glial cells. The molecular mechanisms that senescent astrocytes and microglia exploit to promote the pathological conversion of tau into NFTs within neurons require additional investigation. The absence of neurodegeneration in mice treated with AP (Fig. 3) demonstrates that the attenuation of disease severity does not result from the clearance of neurons containing NFTs. However, it is important to leave open the possibility that other models of neurodegenerative disease may exhibit senescence-associated alterations in cell types other than those observed in the present study. Regardless, based on our observations, it is likely that intervening with senescent-cell accumulation in these models would also reduce the

severity of disease. As this study was designed to prevent senescent cells from accumulating to determine how this affects disease, future studies of senolysis in established disease models will be necessary to determine whether senolytic strategies could translate into the clinic to halt or perhaps revert disease. As senescent cells exhibit a unique and identifiable senescence-associated secretory phenotype, exploiting this phenotype may serve as a possible therapeutic avenue to attenuate many tau-dependent pathologies. Our observation that $p16^{Ink4a}$ expression increases before the onset of NFT aggregation further supports the now commonly held belief that early intervention in these diseases is essential to provide more beneficial effects for patients.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0543-y.

Received: 3 November 2017; Accepted: 29 August 2018;
Published online 19 September 2018.

1. Hayflick, L. & Moorhead, P. S. The serial cultivation of human diploid cell strains. *Exp. Cell Res*. **25**, 585–621 (1961).
2. Coppé, J. P. et al. Senescence-associated secretory phenotypes reveal cell-nonautonomous functions of oncogenic RAS and the p53 tumor suppressor. *PLoS Biol*. **6**, e301 (2008).
3. Baker, D. J. et al. Naturally occurring $p16^{Ink4a}$-positive cells shorten healthy lifespan. *Nature* **530**, 184–189 (2016).
4. Baker, D. J. et al. Clearance of $p16^{Ink4a}$-positive senescent cells delays ageing-associated disorders. *Nature* **479**, 232–236 (2011).
5. Childs, B. G. et al. Senescent intimal foam cells are deleterious at all stages of atherosclerosis. *Science* **354**, 472–477 (2016).
6. Jeon, O. H. et al. Local clearance of senescent cells attenuates the development of post-traumatic osteoarthritis and creates a pro-regenerative environment. *Nat. Med*. **23**, 775–781 (2017).
7. Bhat, R. et al. Astrocyte senescence as a component of Alzheimer's disease. *PLoS ONE* **7**, e45069 (2012).
8. Tan, F. C., Hutchison, E. R., Eitan, E. & Mattson, M. P. Are there roles for brain cell senescence in aging and neurodegenerative disorders? *Biogerontology* **15**, 643–660 (2014).
9. Luo, X. G., Ding, J. Q. & Chen, S. D. Microglia in the aging brain: relevance to neurodegeneration. *Mol. Neurodegener*. **5**, 12 (2010).
10. Yoshiyama, Y. et al. Synapse loss and microglial activation precede tangles in a P301S tauopathy mouse model. *Neuron* **53**, 337–351 (2007).
11. Childs, B. G. et al. Senescent cells: an emerging target for diseases of ageing. *Nat. Rev. Drug Discov*. **16**, 718–735 (2017).
12. Dimri, G. P. et al. A biomarker that identifies senescent human cells in culture and in aging skin *in vivo*. *Proc. Natl Acad. Sci. USA* **92**, 9363–9367 (1995).
13. Stollewerk, A., Klămbt, C. & Cantera, R. Electron microscopic analysis of *Drosophila* midline glia during embryogenesis and larval development using beta-galactosidase expression as endogenous cell marker. *Microsc. Res. Tech*. **35**, 294–306 (1996).
14. Ming, G. L. & Song, H. Adult neurogenesis in the mammalian brain: significant answers and significant questions. *Neuron* **70**, 687–702 (2011).
15. Buenz, E. J. et al. Apoptosis of hippocampal pyramidal neurons is virus independent in a mouse model of acute neurovirulent picornavirus infection. *Am. J. Pathol*. **175**, 668–684 (2009).
16. Chang, J. et al. Clearance of senescent cells by ABT263 rejuvenates aged hematopoietic stem cells in mice. *Nat. Med*. **22**, 78–83 (2016).
17. Zhu, Y. et al. Identification of a novel senolytic agent, navitoclax, targeting the Bcl-2 family of anti-apoptotic factors. *Aging Cell* **15**, 428–435 (2016).
18. Karpel-Massler, G. et al. Induction of synthetic lethality in IDH1-mutated gliomas through inhibition of Bcl-xL. *Nat. Commun*. **8**, 1067 (2017).
19. Flanary, B. E., Sammons, N. W., Nguyen, C., Walker, D. & Streit, W. J. Evidence that aging and amyloid promote microglial cell senescence. *Rejuvenation Res*. **10**, 61–74 (2007).
20. Salminen, A. et al. Astrocytes in the aging brain express characteristics of senescence-associated secretory phenotype. *Eur. J. Neurosci*. **34**, 3–11 (2011).
21. Streit, W. J., Braak, H., Xue, Q. S. & Bechmann, I. Dystrophic (senescent) rather than activated microglial cells are associated with tau pathology and likely precede neurodegeneration in Alzheimer's disease. *Acta Neuropathol*. **118**, 475–485 (2009).
22. Chinta, S. J. et al. Cellular senescence is induced by the environmental neurotoxin paraquat and contributes to neuropathology linked to Parkinson's disease. *Cell Rep*. **22**, 930–940 (2018).

Acknowledgements The authors thank C. H. Cho for numerous contributions to the experiments; the laboratory of C. Howe and specifically M. Standiford for help with the microglia and astrocyte cultures; M. Poeschla for assistance with the phospho-tau immunohistochemistry; G. Nelson for genotyping and animal support; B. Childs for input and assistance in Gal-TEM; the Mayo Clinic Microscopy and Cell Analysis Core and staff for assistance with flow cytometry and transmission electron microscopy; the Mayo Clinic Medical Genome Facility Gene Expression Core for RT–qPCR instrumentation; and R. Petersen and C. Howe for feedback on the manuscript. This work was supported by the Ellison Medical Foundation, the Glenn Foundation for Medical Research, the National Institutes of Health (R01AG053229), the Mayo Clinic Children's Research Center and the Alzheimer's Disease Research Center of Mayo Clinic (all to D.J.B.).

Reviewer information *Nature* thanks M. Serrano and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.J.B. and A.A. performed most of the experiments. C.F.M. assisted with the senescent cell identification by Gal-TEM and FACS. B.L.S. performed immunohistochemistry assessments. J.M.v.D. assisted with experimental design and data interpretation. The manuscript was written by T.J.B. and D.J.B. All authors discussed results, made figures and edited the manuscript. D.J.B. conceived, directed and supervised all aspects of the study.

Competing interests D.J.B. and J.M.v.D. are co-inventors on patent applications licensed to or filed by Unity Biotechnology, a company developing senolytic medicines, including small molecules that selectively eliminate senescent cells. J.M.v.D. is a co-founder of Unity Biotechnology. Research in the Baker laboratory has been reviewed by the Mayo Clinic Conflict of Interest Review Board and is being conducted in compliance with Mayo Clinic Conflict of Interest policies.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0543-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0543-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.J.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mouse strains and drug treatment. $MAPT^{P301S}PS19$ (*PS19*) mice were purchased from The Jackson Laboratory (stock no. 008169) and bred to C57BL/6 for three generations. C57BL/6 *ATTAC* transgenic mice are as previously described[3,4]. Male *PS19* mice were bred to *ATTAC* females to generate cohorts of *ATTAC* and *PS19*;*ATTAC* mice. All mice were on a pure C57BL/6 genetic background. Mice from this cohort were randomly assigned to receive AP20187 (AP; B/B homodimerizer; Clontech) or vehicle twice a week beginning at weaning age (3 weeks). Dosing of AP was 2.0 mg per kg body weight. Six-month-old short-term AP-pulse-treated mice (Extended Data Fig. 6a) received a dose of 10 mg per kg body weight for five consecutive days before tissue collection. Senolytic intervention was performed in C57BL/6 wild-type and *PS19* mice. At weaning, mice were assigned to receive either ABT263 (Cayman, 923564-51-6) or vehicle (Phosal 50 PG, Lipoid NC0130871, 60%; PEG400, Sigma 91893, 30%; ethanol, 10%). ABT263 was administered by oral gavage at a dose of 50 mg per kg body weight on a repeating regimen of five consecutive days of treatment followed by 16 days of rest. Mice were housed in a 12 h:12 h light:dark cycle environment in pathogen-free barrier conditions as previously described in detail[3]. Compliance with relevant ethical regulations and all animal procedures were reviewed and approved by the Mayo Clinic Institutional Animal Care and Use Committee.

Statistical analysis. Prism software was used for all statistical analysis. A Student's two-tailed unpaired *t*-test with Welch's correction was used in Fig. 1a and Extended Data Fig. 4b–e; two-way ANOVA with Tukey's multiple comparisons test was used for Fig. 3d and Extended Data Fig. 10; and one-way ANOVA with Tukey's multiple comparisons test was used in all other figures. For consistency in these comparisons, the following denotes significance in all figures: $*P<0.05$, $**P<0.01$, $***P<0.001$. We note that no power calculations were used. Sample sizes are based on previously published experiments in which differences were observed. No samples were excluded. Investigators were blinded to allocation during experiments and outcome assessment, except for rare instances in which blinding was not possible.

Senescence-associated β-galactosidase transmission electron microscopy (Gal-TEM). Detection of X-Gal crystals by TEM after senescence-associated β-galactosidase (SA-β-Gal) staining was performed as previously described[3,5], with the following alterations to accommodate central nervous tissue. Mice were transcardially perfused with ice-cold Dulbecco's phosphate-buffered saline (DPBS; pH 7.4) until fluid run-off was clear. This was followed by perfusion with 4% paraformaldehyde (PFA) for 10 min at a rate of 3 ml min^{-1}, and then ice-cold DPBS was perfused again for 2 min at the same rate to remove the remaining fixative. Brains were then isolated and the hippocampus and cortex were dissected out. A 1 mm × 1 mm piece from the CA1 and M1 regions, respectively, was then incubated in SA-β-Gal staining solution (Cell Signaling) at 37 °C for 6 h (hippocampus) or 18 h (cortex). The samples were placed in Trump's fixative overnight at 4 °C before being processed for routine transmission electron microscopy (dehydration by xylene-alcohol series, osmium tetroxide staining, and Epon resin embedding). Images were acquired and quantified using a Jeol 1400+ electron microscope with an 80-kV acceleration voltage. Two grids from each tissue were produced, and >100 cells were scanned per grid at a magnification of 20,000× to detect cells containing X-Gal crystals. On average, half of all cells examined were neurons. Cells with one or more crystals, and the total number of cells, were counted. Cells containing crystals were imaged and independently assessed for distinguishing morphology. To define cell type, the following criteria were applied. Astrocytes: circular nucleus with spattered electron density pattern; microglia: abnormally shaped nucleus with a much darker, often phagosome-containing cytoplasm; and neuron: large circular nucleus with less electron density and periodically denoted by an offshooting axon. Only cells with morphology consistent with astrocytes or microglia were clearly positive for X-gal crystals.

Western blotting for soluble and sarkosyl insoluble proteins. Half brains were weighed and homogenized in 5× volume of Buffer I (50 mM Tris base (pH 7.4), 50 mM NaCl, 1 mM EDTA, 1 mM phenylmethylsulfonyl fluoride (PMSF), 1× Halt Protease and Phosphatase Inhibitor Cocktail (Thermo)). Two hundred and fifty microlitres of the homogenate was then added to an equal volume of Buffer S (50 mM Tris (pH 8.0), 274 mM NaCl, 5 mM KCl, 1 mM PMSF, 1X Halt Protease and Phosphatase Inhibitor Cocktail (Thermo)) and ultracentrifuged at 150,000*g* for 15 min at 4 °C. The supernatant (S1-soluble protein fraction) was transferred to a new tube and the pellet homogenized in 3× volume of sucrose buffer (10 mM Tris (pH 7.4), 0.8M NaCl, 10% sucrose, 1 mM EGTA, 1 mM PMSF) before being ultracentrifuged at 150,000*g* for 15 min at 4 °C. The pellet was discarded and the supernatant incubated with sarkosyl (sodium lauroyl sarcosinate) at a final concentration of 1% for 1 h at 37 °C. After incubation, the samples were ultracentrifuged at 150,000g for 30 min at 4 °C. The supernatant was discarded and the pellet was resuspended in 25 μl Buffer F (10 mM Tris (pH 8.0), 1 mM EDTA) to obtain the insoluble protein fraction (S2). Equal parts of 2× laemmli buffer (Bio-Rad) containing 5% β-mercaptoethanol was added to each fraction (S1 and S2) and boiled at 100 °C for 15 min to prepare the sarkosyl-soluble and -insoluble protein lysates. For total protein lysate, 90 μl of the homogenate (half brain in 5X volume Buffer I) was added to 110 μl of Buffer T (2% SDS, 50 mM Tris (pH 7.4), 274 mM NaCl, 5 mM KCl, 5 mM EDTA, 1% Triton-X-100, 1 mM PMSF, X Halt Protease and Phosphatase Inhibitor Cocktail (Thermo)). The samples were then sonicated and centrifuged at 16,000*g* for 15 min at 4 °C to remove debris. The supernatant was removed and added to equal parts 2× laemmli buffer with 5% β-mercaptoethanol and boiled at 100 °C for 15 min to prepare the total protein lysate. Western blotting was performed as previously described[23]. Blots were probed with antibodies for total tau (Thermo Fisher; MN1000, 1:5,000) and phospho-tau S202/T205 (Thermo Fisher; MN1020, 1:1,000). Ponceau-S staining was performed to normalize lysate loading for the total- and S1-fraction lysates. Quantification was performed using ImageJ as previously described[3].

Quantitative RT–PCR. RNA extraction, cDNA synthesis and RT–qPCR analysis were performed on hippocampi and cortical samples from mouse brains as previously described[24]. Primers used to amplify *Casp8*, *GFP*, $p16^{Ink4a}$, $p19^{Arf}$, *p21*, *Pai1*, *Il-6*, *Il-1b* and *Cd11b* were as previously described[3,5,24]. The following additional primers were used: *Gfap* forward 5′-CCTTCTGACACGGATTTGGT-3′, reverse 5′-TAAGCTAGCCCTGGACATCG-3′; *S100b* forward 5′-CCGGAGTACTGGTGGAAGAC-3′, reverse 5′-GGACACTGAAGCCAGAGAGG-3′; *Aqp4* forward 5′-TGAGCTCCACATCAGGACAG-3′, reverse 5′-TCCAGCTCGATCTTTTGGAC-3′; *Cx3cr1* forward 5′-GTTCCAAAGGCCACAATGTC-3′, reverse 5′-TGAGTGACTGGCACTTCCTG-3′; *Olig2* forward 5′-CCCCAGGATGATCTAAGC-3′, reverse 5′-CAGAGCCAGGTTCTCCTCC-3′; *Nefl* forward 5′-AGGCCATCTTGACATTGAGG-3′, reverse 5′-GCAGAATGCAGACATTAGCG-3′; *Tbp* forward 5′-GGCCTCTCAGAAGCATCACTA-3′, reverse 5′-GCCAAGCCCTGAGCATAA-3′. Expression for all experiments was normalized first to *Tbp*.

Immunohistochemistry and immunofluorescence staining. Mice were transcardially perfused as described in 'Senescence-associated β-galactosidase transmission electron microscopy (Gal-TEM)'. Brains were stored in 4% PFA overnight at 4 °C and then cryoprotected by incubating in a 30% sucrose solution for 48 h at 4 °C. Samples were sectioned into 30-μM-thick coronal sections and stored in antifreeze solution (300 g sucrose, 300 ml ethylene glycol, 500 ml PBS) at −20 °C. Nissl staining (bregma −2.1 to −2.4 mm), thioflavin S staining (bregma −1.4 to −1.6 mm), and phospho-tau S202/S205 (Thermo Fisher, MN1020; 1:500), phospho-tau T231 (Thermo Fisher, MN1040; 1:500), phospho-tau S396 (Abcam, 109390; 1:500), and GFAP (Dako, Z0334; 1:500) and IBA1 (Novus, NB100-1028; 1:100) immunohistochemistry staining (bregma 1.6 to 1.0 mm and lateral 2.0 to 2.7 mm) was performed on free-floating sections as described[25–27]. NeuN staining (EMD, MAB377; 1:200) of five sections (between bregma −1.3 to −2.5) to measure the dentate gyrus area was performed as previously described[28]. For cellular proliferation assays, mice were injected with EdU (Carbosynth, NE08701; 75 mg kg^{-1}) intraperitoneally 24 h before euthanasia. Imaging of EdU-positive cells (lateral 0.75 to 1.25 mm) was performed following the manufacturer's instructions (Invitrogen Click-iT EdU Alexa Fluor 488 Imaging Kit, C10337). IBA1 (Wako, 019-19741; 1:500) immunofluorescent staining and IBA1/EdU colocalization assessments were performed as previously described[28]. TUNEL staining (lateral 0.75 to 1.25 mm) was performed according to the manufacturer's instructions (Roche In situ Cell Death Detection Kit, Fluorescein: 11684795910). Thioflavin S, EdU/IBA1 colocalization, and in vivo TUNEL-stained images were acquired on a Zeiss LSM 780 confocal system using multi-track configuration.

Single-cell preparation and FACS. Dissociation of cerebral tissue was performed using the Adult Brain Dissociation kit from Miltenyi (MACS, 130-107-677), according to the manufacturer's instructions. Samples were then incubated with a viability dye, LIVE/DEAD Aqua (Invitrogen, L34966; 1:250) followed by incubation with CD11b eFluor 450 (eBioscience, 48-0112-80, 1:100), CD45 APC eFluor 780 (eBioscience, 47-0451-82; 1:200), Glast1 PE (Miltenyi Biotec, 130-095-821; 1:100), O1 AF 700 (R&D Systems, FAB1327N-100UG; 1:100) and CD56 APC (R&D Systems, FAB7820A; 1:100). These samples were then sorted using a FACSAria IIu SORP (BD Biosciences), with gating parameters created using FACSDiva 8.0.1 (BD Biosciences). A precise gating strategy was used to maximize the purification of each isolated cell population. In brief, populations were isolated first by a negative report of the viability dye indicating the cell is viable (Extended Data Fig. 4), followed by a positive report of the desired marker, then negative reports of the other labels used. This strategy allowed for live cells containing only the desired marker to be sorted, while eliminating dead cells. Cells were sorted directly into lysis buffer and RNA was extracted with RNeasy Micro kits according to the manufacturer's instructions (Qiagen, cat no. 74004). cDNA synthesis and RT–qPCR analysis were performed as described above.

Novel-object recognition. Novel-object recognition testing was performed as previously described[15]. In brief, mice from each cohort were acclimatized to a 50 cm × 50 cm testing environment for a period of two minutes. After acclimatization, the mice were removed, the testing area was cleaned with 70% ethanol, and two

identical scented candles were placed in either corner of the testing area approximately 5 cm from either wall. Mice were reintroduced, and the ratio of both the number of visits and the time spent at each candle was recorded for a period of ten minutes. Recording was performed from above (Panasonic WV-CP294) and all video files were analysed with TopScan Version 3.00 (Clever Sys). Afterwards, the mice were removed, the testing area cleaned with 70% ethanol, and one candle was replaced with a novel scent. The mice were reintroduced and the number of visits and total time per candle was recorded as before. Testing was also performed with visual stimuli, by placing identical toy brick towers at either corner and then replacing with a different toy brick tower in the testing phase, using the same experimental paradigm monitoring for the number of investigations.

In vitro astrocyte and microglia culture. Astrocyte and microglia primary cultures were prepared in tandem from mixed glial cultures as previously described[29]. C57BL/6 wild-type and *ATTAC* pups (postnatal day (P)0–P3) were euthanized, and the cerebellum was discarded. Meninges were removed from the remaning tissue using forceps and a dissection scope. Cleaned cerebral tissue was placed in chilled Earle's Balanced Salt Solution with HEPES (EBSH) (NaCl 120 mM, NaH_2PO_4 10 mM, KCl 2.5 mM, $C_6H_{12}O_6$ 20 mM, HEPES 20 mM, $NaHCO_3$ 10 mM, BSA 0.3%, H_2O) until the remaining mice were euthanized, and then mice were pooled together on the basis of genotype (3–4 brains per group). The tissue was minced using a razor blade and dissociated by shaking in a 0.025% Trypsin/EBSH solution at 37 °C for 30 min. FBS and $MgSO_4$ (3.82%)/ DNase I (1 mg ml^{-1}) were added, and the sample was placed on ice for 5 min to halt trypsinization. Samples were mixed and centrifuged at 200*g* at 25 °C for 5 min. The supernatant was discarded, and the remaining pellet was resuspended in EBSH. Tissue was triturated using a 1 ml pipette to completely dissociate the sample and allowed to settle for 5 min to remove large debris. Samples were then transferred to clean tubes and underlaid with a 4% BSA/EBSH solution. The tissue was then centrifuged at 100*g* at 25 °C for 8 min. Cells were counted using trypan blue and a haemocytometer and plated on a poly-D-lysine-coated T75 dish (7–10 million cells per flask) with glial cell culture medium (GCM) consisting of DMEM with 10% FBS, sodium pyruvate (1 mM), Pen/Strep (500 μg ml^{-1}), and InvivoGen Primocin (100 μg ml^{-1}). Cultures were grown for 14 days (37 °C, ambient O_2) with medium changes every 4 days. Microglia were isolated as previously described[30] using the EasySep Mouse CD11b Positive Selection Kit from Stem Cell (cat. no. 18970). Microglia were collected and plated on 10-well glass slides (5,000 cells per well) and cultured for 6 days in GCM with LADMAC-conditioned medium (20%, provided by the Howe laboratory) before further experimentation. This conditioned medium aids in the proliferation and maintenance of microglia cultures through the secretion of macrophage colony-stimulating factor by the LADMAC cells[31]. Microglia were allowed to proliferate for 6 days before experimentation. The mixed glial culture flow-through from the EasySep CD11b kit was replated in GCM on a poly-D-lysine-coated T75 dish (10 million cells per flask). These cultures then underwent purification for astrocytes as previously described[29]. After 48 h, flasks were placed on an orbital shaker and agitated at 200 r.p.m. for two 24-h periods with medium refreshed once during and after the shaking. Flasks were then exposed to GCM containing liposomal clodronate (Clodrosome, 8909; 100 μg ml^{-1}) for 72 h to remove any remaining microglia from the culture. The liposomal clodronate medium was then removed and culture plates washed before further experimentation.

Microglia activation and TUNEL staining. Microglia samples were exposed to medium containing IFNγ (R&D Systems, 285-IF; 200 ng ml^{-1}), lipopolysaccharides (LPS) (Sigma, L2654; 100 ng ml^{-1}) or a combination of both for a period of 24 h to induce an inflammatory response[32]. Cells were then processed for immunofluorescence to determine inflammation state as previously described[33]. Anti-CD11b antibody (Bio-Rad, MCA711G; 1:500) and goat anti-rat AlexaFluor 594 (Invitrogen, A-11007; 1:500) staining was counterstained with DAPI (Invitrogen, D1306; 1:1,000). To assess AP-mediated cell clearance specificity, activated or basal microglia were exposed to AP20187 (Clontech, 635059; 10 nM or 100 nM) for a period of 24 h. TUNEL staining was then performed according to the manufacturer's instructions (Roche In situ Cell Death Detection Kit, Fluorescein: 11684795910). All imaging was performed using an Olympus BX53 Fluorescence microscope and DP80 digital camera. Analysis was performed using the Fiji distribution of ImageJ (version 1.51n)[34]. To obtain a TUNEL positive percentage, a region of interest was defined using Li Auto Thresholding of the DAPI channel, and the colocalization percentage was calculated using the colocalization threshold plugin bound by that region.

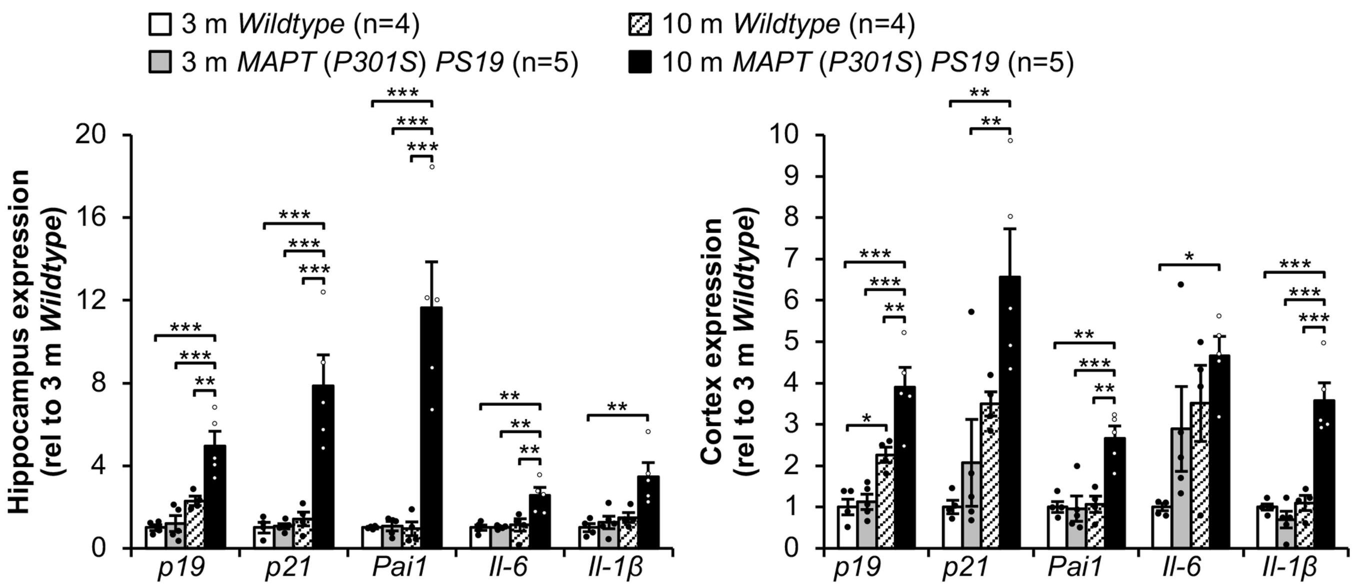
IncuCyte tracking of basal and activated astrocytes. To track basal and activated astrocytic response to AP, astrocytes were plated in a 48-well culture plate (10,000 cells per well) and placed into the IncuCyte S3 Live-Cell Analysis System. The IncuCyte System is a time-lapse imaging system that records changes in cell culture through photographic capture of the culture well within the incubator. Cultures were acclimatized to the system for a period of 6 h, then exposed to medium containing IFNγ (R&D Systems, 285-IF; 200 ng ml^{-1}), LPS (Sigma, L2654; 100 ng ml^{-1}) or a combination of both for a period of 24 h to induce an inflammatory response[35]. Cells were also plated on 10-well slides and processed in tandem for immunofluorescence staining to verify activation status with anti-GFAP (DAKO, Z0334; 1:500) and counterstained with DAPI (Invitrogen, D1306; 1:1,000). After activation, astrocytes were exposed to AP20187 (Clontech, 635059; 10 nM or 100 nM) for a period of 24 h. The IncuCyte-captured phase images of each culture well were taken every 30 min over this period using the following settings: segmentation adjustment: 0.8, hole fill: 450, adjust size (pixels): −1, Minimum area (μm^2): 0.1. The phase confluency difference was calculated by subtracting the final phase confluency of each image from its initial value.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

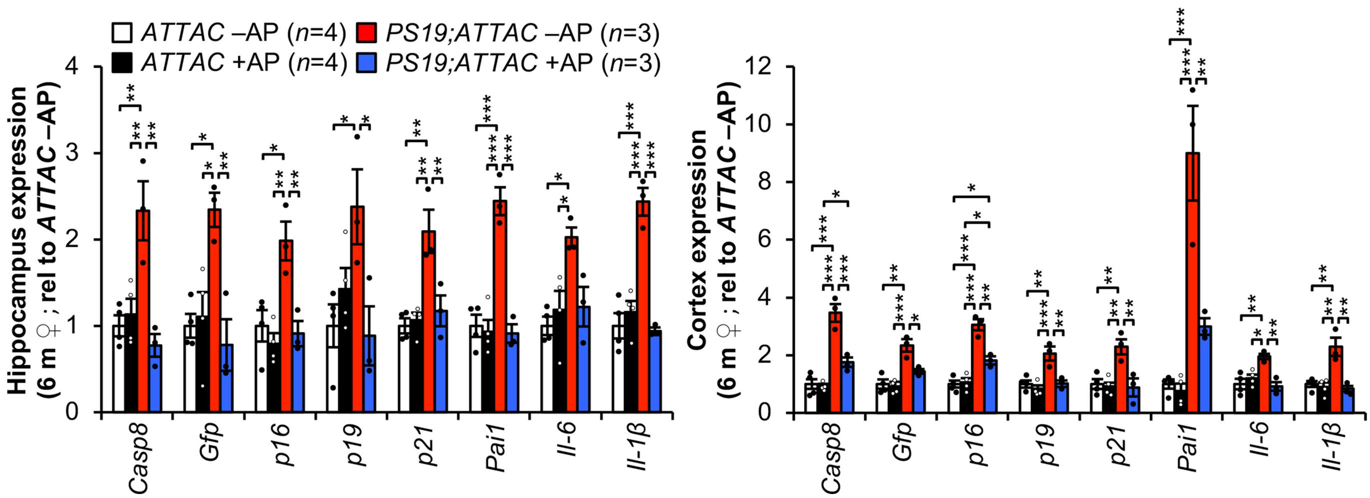
Data availability

All source data and exact *P* values (if applicable) for every figure are included in the supporting information that accompanies the paper.

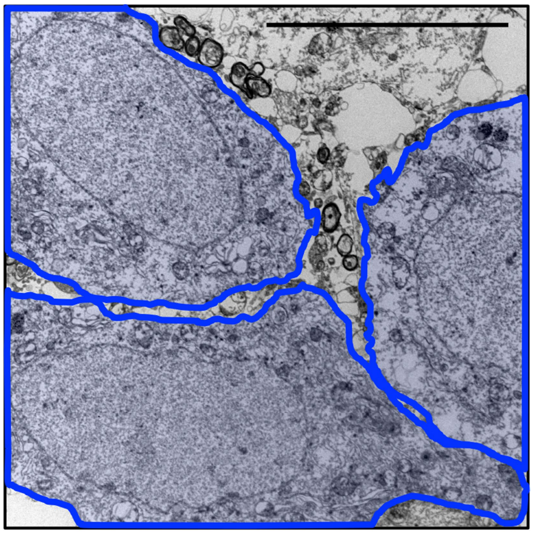
23. Kasper, L. H. et al. CREB binding protein interacts with nucleoporin-specific FG repeats that activate transcription and mediate NUP98-HOXA9 oncogenicity. *Mol. Cell. Biol*. **19**, 764–776 (1999).
24. Baker, D. J. et al. Opposing roles for p16^{Ink4a} and p19^{Arf} in senescence and ageing caused by BubR1 insufficiency. *Nat. Cell Biol*. **10**, 825–836 (2008).
25. Parent, J. M., von dem Bussche, N. & Lowenstein, D. H. Prolonged seizures recruit caudal subventricular zone glial progenitors into the injured hippocampus. *Hippocampus* **16**, 321–328 (2006).
26. Ly, P. T., Cai, F. & Song, W. Detection of neuritic plaques in Alzheimer's disease mouse model. *J. Vis. Exp*. **53**, 2831 (2011).
27. Oh, K. J. et al. Staging of Alzheimer's pathology in triple transgenic mice: a light and electron microscopic analysis. *Int. J. Alzheimers Dis*. **2010**, 780102 (2010).
28. Yang, Z. et al. Age-related decline in BubR1 impairs adult hippocampal neurogenesis. *Aging Cell* **16**, 598–601 (2017).
29. Kumamaru, H. et al. Liposomal clodronate selectively eliminates microglia from primary astrocyte cultures. *J. Neuroinflammation* **9**, 116 (2012).
30. Gordon, R. et al. A simple magnetic separation method for high-yield isolation of pure primary microglia. *J. Neurosci. Methods* **194**, 287–296 (2011).
31. Tsay, H. J. et al. Amyloid β peptide-mediated neurotoxicity is attenuated by the proliferating microglia more potently than by the quiescent phenotype. *J. Biomed. Sci*. **20**, 78 (2013).
32. Chao, C. C., Hu, S., Molitor, T. W., Shaskan, E. G. & Peterson, P. K. Activated microglia mediate neuronal cell injury via a nitric oxide mechanism. *J. Immunol*. **149**, 2736–2741 (1992).
33. Wang, G. et al. Apoptosis and proinflammatory cytokine responses of primary mouse microglia and astrocytes induced by human H1N1 and avian H5N1 influenza viruses. *Cell. Mol. Immunol*. **5**, 113–120 (2008).
34. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
35. Chung, I. Y. & Benveniste, E. N. Tumor necrosis factor-alpha production by astrocytes. Induction by lipopolysaccharide, IFN-gamma, and IL-1 beta. *J. Immunol*. **144**, 2999–3007 (1990).



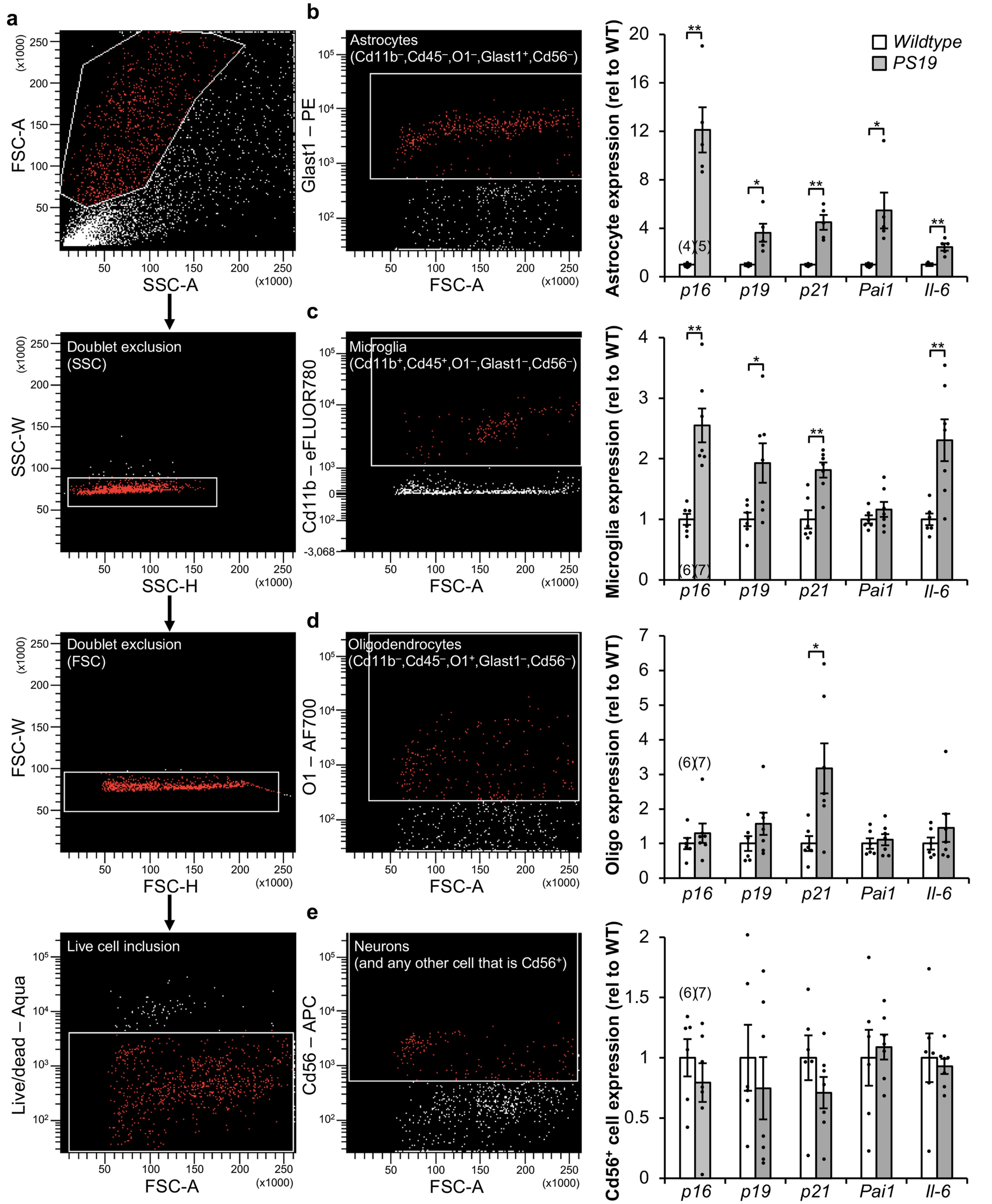
Extended Data Fig. 1 | Senescent cells accumulate in *PS19* mice. RT–qPCR analysis of senescence-associated genes in hippocampi (left) and cortices (right) of three- and ten-month-old male mice. Number of mice is as indicated, two independent experiments; normalized to the three-month wild-type group. Data are mean ± s.e.m. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$ (one-way ANOVA with Tukey's multiple comparisons test). Exact P values can be found in the accompanying Source Data.



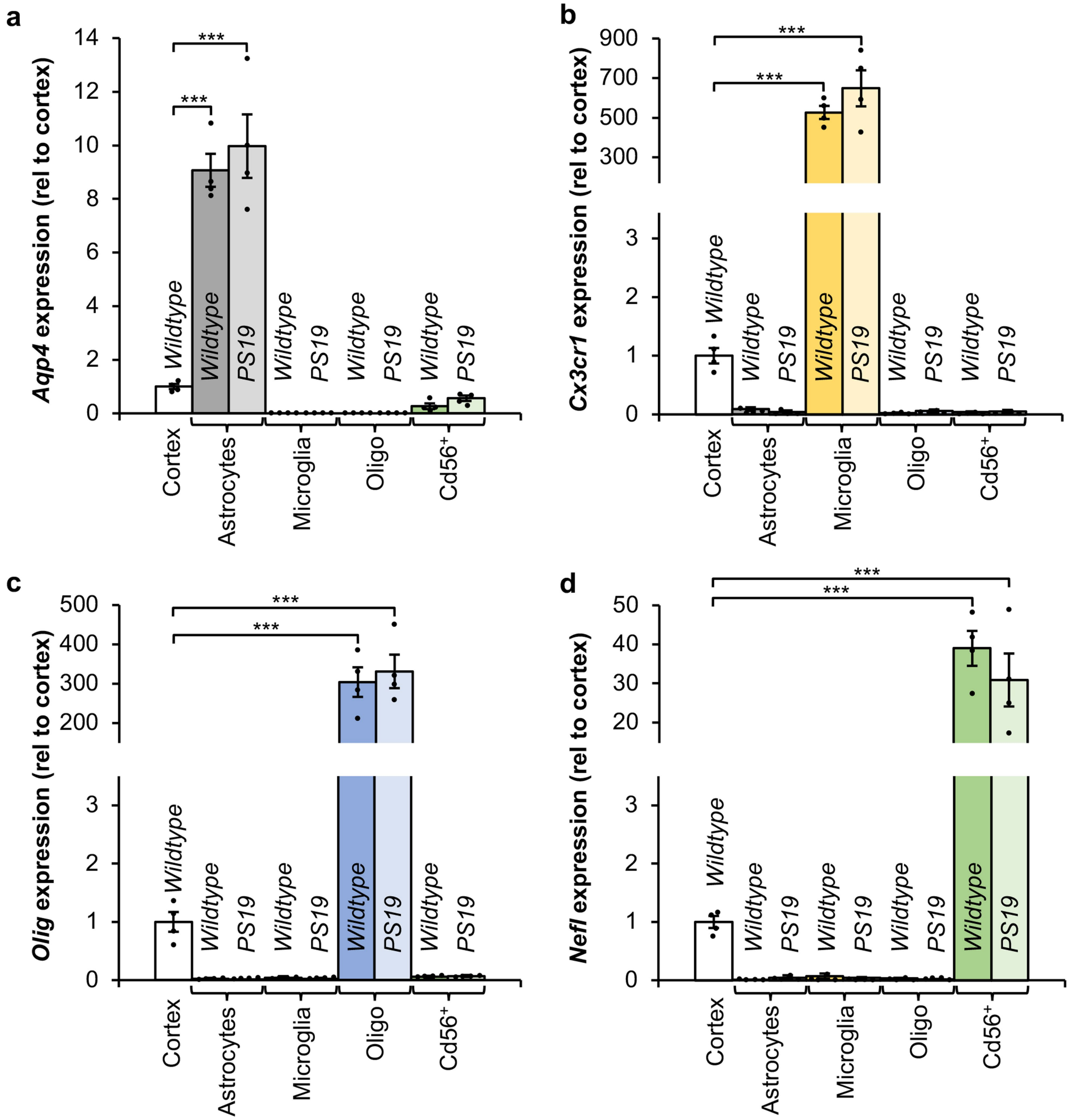
Extended Data Fig. 2 | AP-mediated clearance selectively removes senescent cells that accumulate in the brains of *PS19;ATTAC* mice. RT–qPCR analysis of the expression of senescence markers in the hippocampus (left) and cortex (right) of six-month-old female mice, treated with either vehicle (–AP) or AP20187 (+AP). Number of mice is as indicated; normalized to the *ATTAC* – AP group. Data are mean ± s.e.m. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$ (one-way ANOVA with Tukey's multiple comparisons test). Exact *P* values can be found in the accompanying Source Data.



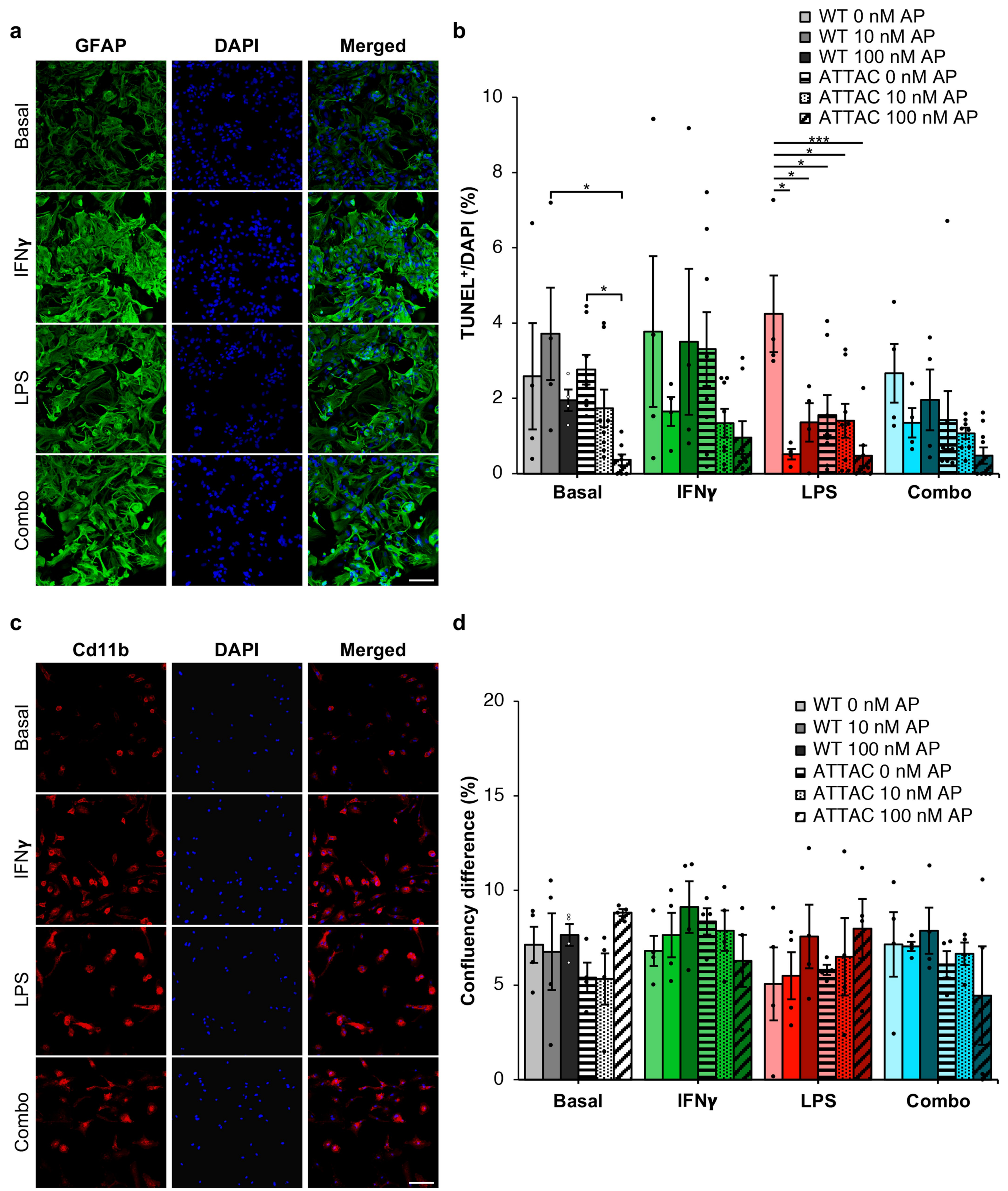
Extended Data Fig. 3 | Neurons do not exhibit X-Gal crystals upon Gal-TEM imaging. Representative electron microscopy image of neurons after SA-β-Gal staining from a six-month-old, vehicle-treated *PS19*;*ATTAC* male mouse ($n = 3$ male mice, 2 independent experiments). The image has been artificially coloured to show individual cell bodies. Scale bar, 10 μm.



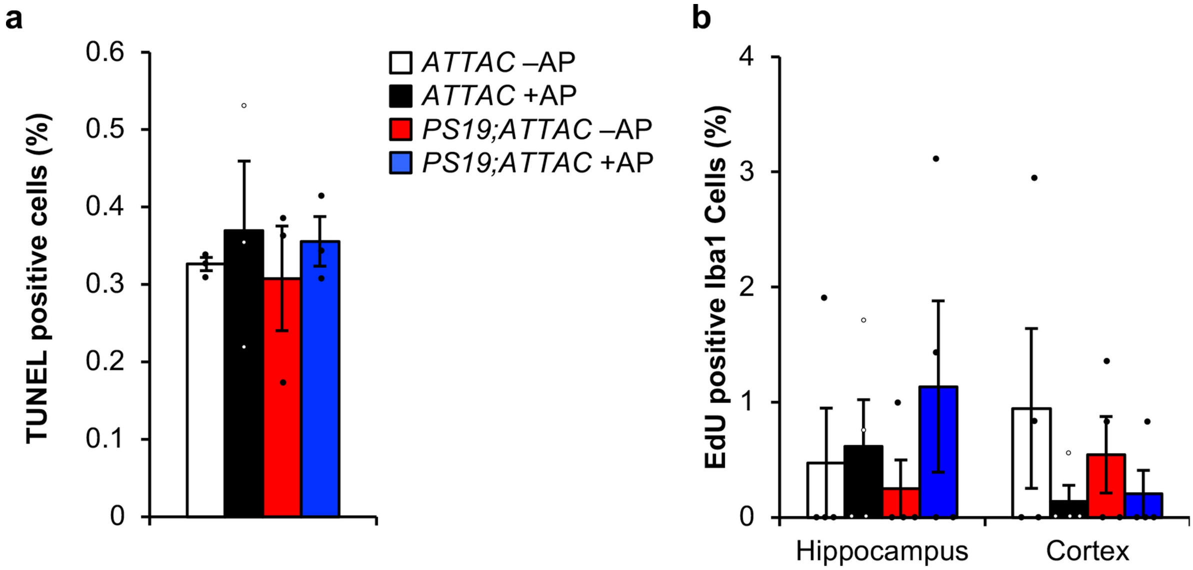
Extended Data Fig. 4 | Increased expression of senescence-associated genes is observed in astrocytes and microglia isolated from *PS19* mice. a–e, Gating strategy (**a**) for FACS isolation of living astrocytes (**b**), microglia (**c**), oligodendrocytes (**d**) and neuron-enriched CD56$^+$ cells (**e**) from cortices of six-month-old wild-type and *PS19* mice. **b**, Astrocyte (CD11b$^-$CD45$^-$O1$^-$GLAST$^+$CD56$^-$) fraction (left) and RT–qPCR analysis (right). **c**, Microglia (CD11b$^+$CD45$^+$O1GLAST$^+$CD56$^-$) fraction (left) and RT–qPCR analysis (right). **d**, Oligodendrocyte (CD11b$^-$CD45$^-$O1$^+$GLAST$^-$CD56$^-$) fraction (left) and RT–qPCR analysis (right). **e**, Neuron-enriched CD56$^+$ (CD11b$^-$CD45$^-$O1$^-$GLAST$^-$CD56$^+$) fraction (left) and RT–qPCR analysis (right). Individual numbers of independent mouse cell population isolations are indicated in the parentheses above $p16^{Ink4a}$ columns (2 independent experiments). Data are mean ± s.e.m. $*P < 0.05$; $**P < 0.01$ (unpaired two-sided *t*-tests with Welch's correction). Exact *P* values can be found in the accompanying Source Data.



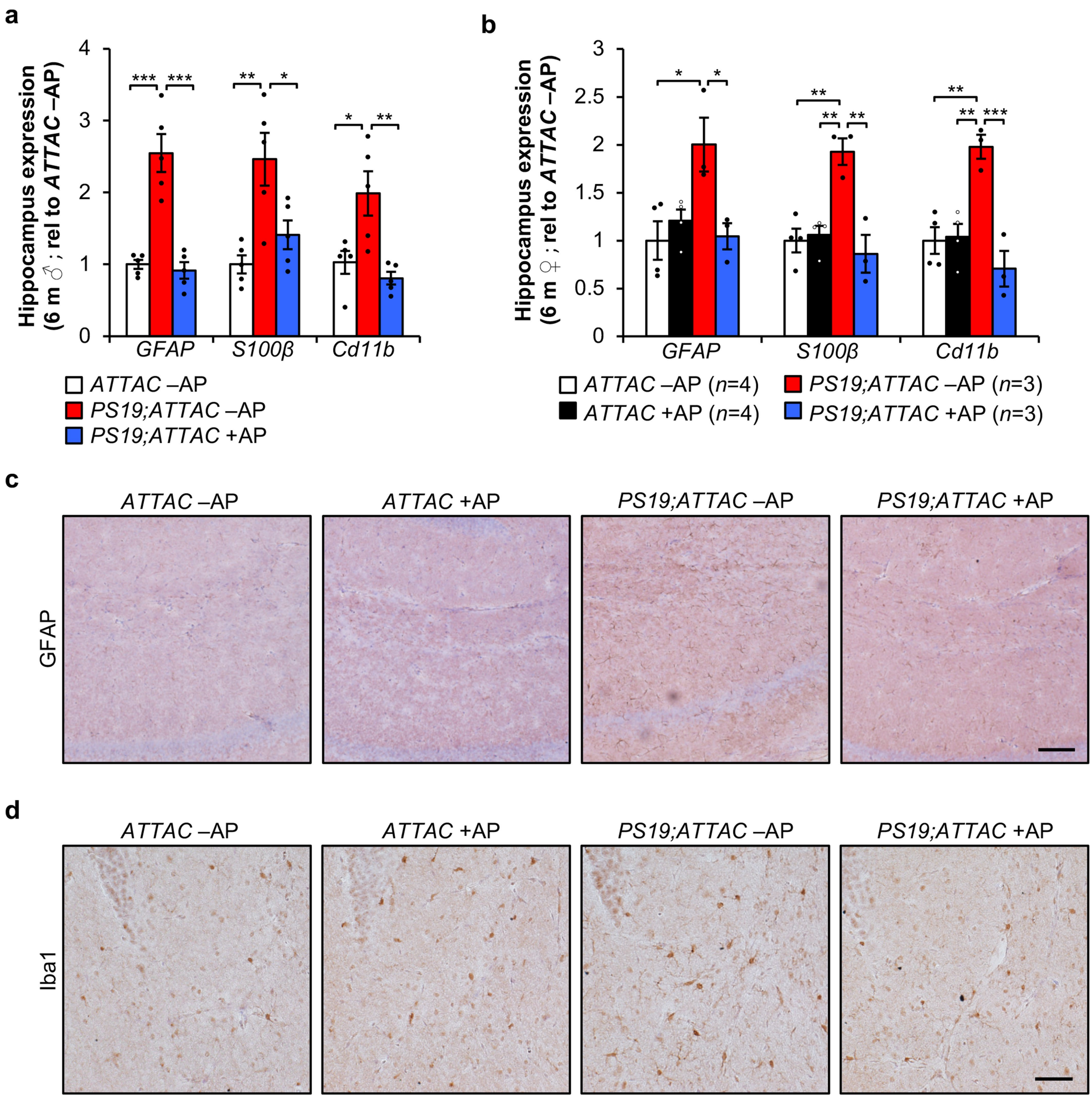
Extended Data Fig. 5 | Verification of the identity of cell populations isolated by FACS. a–d, RT–qPCR analysis of cell-identity markers from cell populations isolated from six-month-old wild-type and *PS19* mice: *Aqp4* expression enriched in astrocytes (**a**), *Cx3cr1* expression enriched in microglia (**b**), *Olig2* expression enriched in oligodendrocytes (**c**) and *Nefl* expression enriched in neurons (**d**). Expression is normalized to intact cortices of six-month-old wild-type mice ($n = 4$ biologically independent cell isolations for each group, 2 independent experiments). Data are mean ± s.e.m. ***$P << 0.001$ (one-way ANOVA with Tukey's multiple comparisons test). Exact P values can be found in the accompanying Source Data.



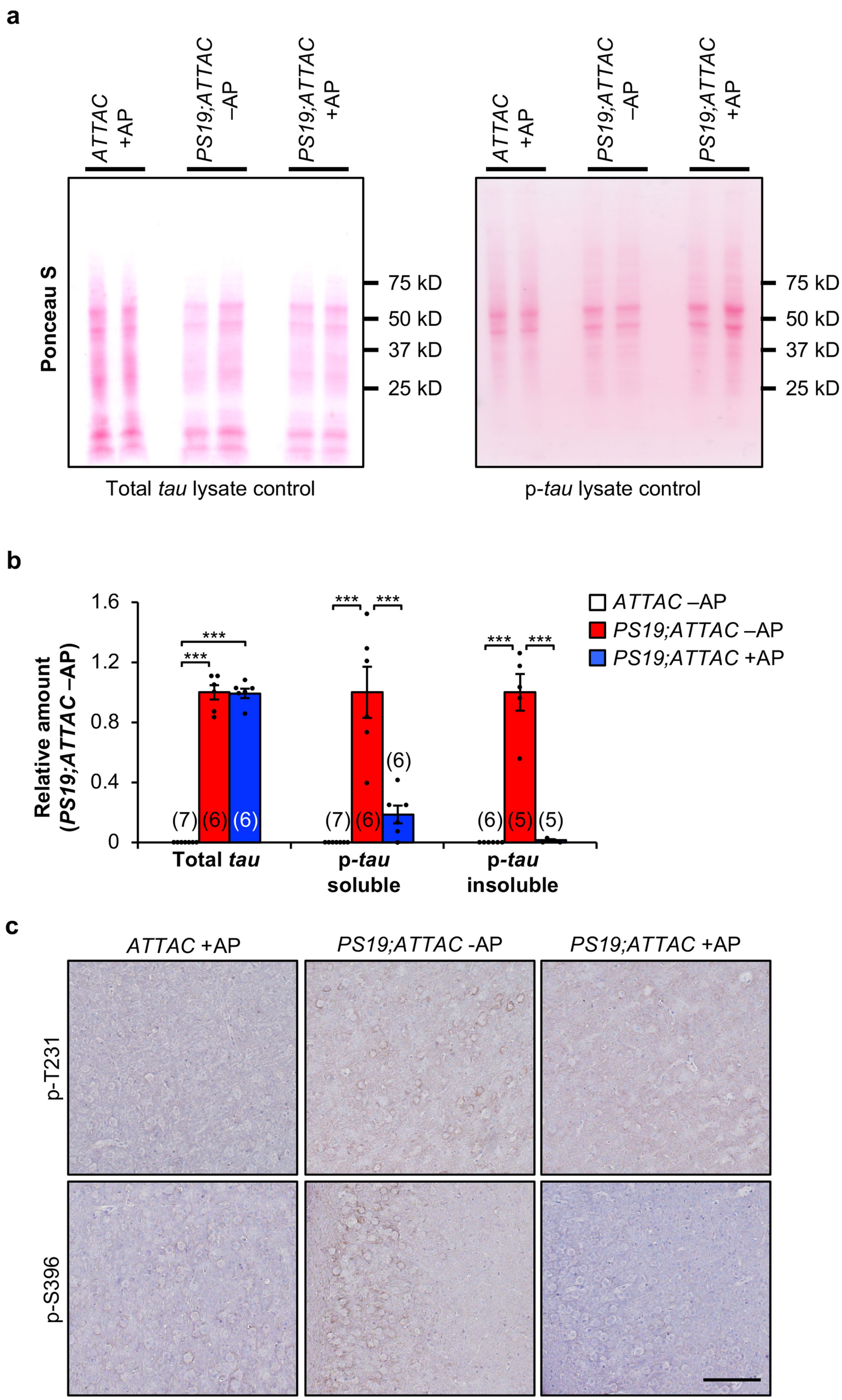
Extended Data Fig. 6 | AP administration does not erroneously eliminate non-senescent glial cells isolated from *ATTAC* mice. a, CD11b staining of primary microglia treated with IFNγ (200 ng ml^{-1}), LPS (100 ng ml^{-1}) or a combination of both ($n = 3$ biologically independent samples). **b**, Quantification of TUNEL-positive bodies in basal or activated microglia ($n = 4$ wild-type and 8 *ATTAC* cultures for each treatment group, 2 independent experiments). **c**, GFAP staining of primary astrocytes treated with IFNγ, LPS or a combination of both as described in **a** ($n = 3$ biologically independent samples). **d**, Quantification of the change in confluency over 24 h in basal or activated astrocytes ($n = 4$ biologically independent cultures of each genotype and treatment). Scale bars, 100 µm (**a**, **c**). Data are mean ± s.e.m. $*P < 0.05$; $***P < 0.001$ (one-way ANOVA with Tukey's multiple comparisons test (**b**, **d**)). Exact P values can be found in the accompanying Source Data.



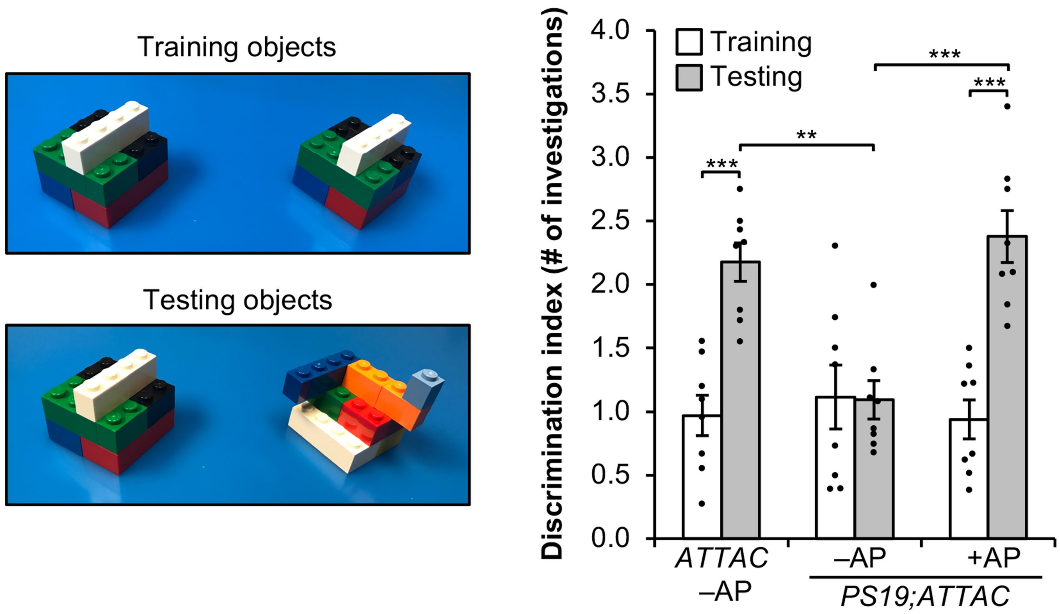
Extended Data Fig. 7 | Administration of AP does not broadly eliminate cells or increase proliferation of microglia. a, Quantification of TUNEL-positive bodies (as a percentage of all cells) at the transition between the CA2 and CA3 within the hippocampus after a short-term AP administration in six-month-old mice ($n = 3$ mice per genotype and treatment group). **b**, Quantification of IBA1/EdU double-positive cells in the hippocampus and cortex of six-month-old mice that were administered AP beginning at weaning age ($n = 4$ mice per genotype and treatment group). Data are mean ± s.e.m. We note that no comparison is statistically significant (one-way ANOVA with Tukey's multiple comparisons test). Exact *P* values can be found in the accompanying Source Data.



Extended Data Fig. 8 | Senescent cells promote gliosis. a, RT–qPCR analysis of *Gfap*, *S100b* and *Cd11b* in the hippocampi of six-month-old male mice ($n = 5$ mice per group; normalized to the *ATTAC* – AP group). **b**, RT–qPCR analysis as in **a** in hippocampi of six-month-old female mice (number of mice as indicated; normalized to the *ATTAC* – AP group). **c**, Representative GFAP immunohistochemistry staining in the hippocampus of six-month-old vehicle and AP-treated *ATTAC* and *PS19*;*ATTAC* female mice ($n = 4$ mice per group, 2 independent experiments). **d**, Representative IBA1 staining in the hippocampus of six-month-old vehicle and AP-treated *ATTAC* and *PS19*;*ATTAC* female mice ($n = 4$ mice per group, 2 independent experiments). Scale bar, 100 μm (**c**) and 50 μm (**d**). Data are mean ± s.e.m. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$ (one-way ANOVA with Tukey's multiple comparisons test). Exact *P* values can be found in the accompanying Source Data.



Extended Data Fig. 9 | AP treatment attenuates tau phosphorylation. **a**, Ponceau-S loading controls for the western-blot lysates of the whole brain of six-month-old mice (shown in Fig. 3a) for total tau (left) and phosphorylated tau (S202/T205; right). **b**, Quantification of the western-blot analysis of the whole brain of six-month-old mice for soluble tau (left), soluble phosphorylated tau (S202/T205; middle), and insoluble phosphorylated tau (S202/T205; right). Numbers of biologically independent mice are indicated in parentheses, data are from ≥3 independent experiments. **c**, Immunostaining of the cortex of six-month-old mice for tau protein phosphorylated at T231 (top) and S396 (bottom; $n = 4$ mice per group, 2 independent experiments). Scale bar, 100 μm. Data are mean ± s.e.m. $^{***}P < 0.001$ (one-way ANOVA with Tukey's multiple comparisons test). Exact P values can be found in the accompanying Source Data.



Extended Data Fig. 10 | Vision-based novel-object discrimination remains intact in AP-treated *PS19;ATTAC* mice. Objects used for novel-object recognition during the training and testing phases for visual discrimination (left) and the average ratio for the number of investigations (right, $n = 8$ female mice per group). Data are mean ± s.e.m. $^{**}P < 0.01$; $^{***}P < 0.001$ (two-way ANOVA with Tukey's multiple comparisons test). Exact P values can be found in the accompanying Source Data.

OPEN
https://doi.org/10.1038/s41586-018-0617-x

Temporal development of the gut microbiome in early childhood from the TEDDY study

Christopher J. Stewart[1,2,18]*, Nadim J. Ajami[1,18], Jacqueline L. O'Brien[1], Diane S. Hutchinson[1], Daniel P. Smith[1], Matthew C. Wong[1], Matthew C. Ross[1], Richard E. Lloyd[1], HarshaVardhan Doddapaneni[3], Ginger A. Metcalf[3], Donna Muzny[3], Richard A. Gibbs[3], Tommi Vatanen[4], Curtis Huttenhower[4], Ramnik J. Xavier[4], Marian Rewers[5], William Hagopian[6], Jorma Toppari[7,8], Anette-G. Ziegler[9,10,11], Jin-Xiong She[12], Beena Akolkar[13], Ake Lernmark[14], Heikki Hyoty[15,16], Kendra Vehik[17], Jeffrey P. Krischer[17] & Joseph F. Petrosino[1]*

The development of the microbiome from infancy to childhood is dependent on a range of factors, with microbial–immune crosstalk during this time thought to be involved in the pathobiology of later life diseases[1–9] such as persistent islet autoimmunity and type 1 diabetes[10–12]. However, to our knowledge, no studies have performed extensive characterization of the microbiome in early life in a large, multi-centre population. Here we analyse longitudinal stool samples from 903 children between 3 and 46 months of age by 16S rRNA gene sequencing ($n = 12{,}005$) and metagenomic sequencing ($n = 10{,}867$), as part of the The Environmental Determinants of Diabetes in the Young (TEDDY) study. We show that the developing gut microbiome undergoes three distinct phases of microbiome progression: a developmental phase (months 3–14), a transitional phase (months 15–30), and a stable phase (months 31–46). Receipt of breast milk, either exclusive or partial, was the most significant factor associated with the microbiome structure. Breastfeeding was associated with higher levels of *Bifidobacterium* species (*B. breve* and *B. bifidum*), and the cessation of breast milk resulted in faster maturation of the gut microbiome, as marked by the phylum Firmicutes. Birth mode was also significantly associated with the microbiome during the developmental phase, driven by higher levels of *Bacteroides* species (particularly *B. fragilis*) in infants delivered vaginally. *Bacteroides* was also associated with increased gut diversity and faster maturation, regardless of the birth mode. Environmental factors including geographical location and household exposures (such as siblings and furry pets) also represented important covariates. A nested case–control analysis revealed subtle associations between microbial taxonomy and the development of islet autoimmunity or type 1 diabetes. These data determine the structural and functional assembly of the microbiome in early life and provide a foundation for targeted mechanistic investigation into the consequences of microbial–immune crosstalk for long-term health.

In this study, a total of 12,500 stool samples from 903 children from three European countries (Germany, Sweden and Finland) and three US states (Colorado, Georgia and Washington) were analysed. The children represent those who seroconverted to islet cell autoantibody positivity or developed type 1 diabetes (T1D) and matched controls. Stool samples were collected, on average, monthly from around 3 months of age as part of the The Environmental Determinants of Type 1 Diabetes in the Young (TEDDY) study[13]. After rarefaction and limiting samples to 3–46 months of age, we analysed the microbiome (16S rRNA gene sequencing, $n = 12{,}005$ samples from 903 children; metagenomic sequencing, $n = 10{,}867$ samples from 783 children) and functional metagenome (metagenomic sequencing only) from longitudinal stool samples (Extended Data Table 1). A companion paper by Vatanen et al.[14] focused exclusively on metagenomic sequencing data.

In this cohort of children that are at-risk for developing islet autoimmunity (IA) or T1D, we aimed to (1) characterize definitively the longitudinal gut microbiome development from 3 to 46 months of age; (2) determine selected maternal and postnatal influences on the developing bacterial community during this same time period of early development; and (3) use a nested case–control analysis to investigate the potential of the microbiome as a predictor for the development of IA or T1D.

A general overview of bacterial taxonomic and functional pathway development is provided in Supplementary Note 1 and Extended Data Fig. 1. Dirichlet multinomial mixtures (DMM) modelling was applied to 16S rRNA gene sequencing (Fig. 1) and metagenomic sequencing data (Extended Data Fig. 2). All samples from 3 to 46 months of age were included, and 16S rRNA gene sequencing profiles formed ten clusters (based on lowest Laplace approximation) (Fig. 1a). Bacterial richness and diversity increased in each cluster (Fig. 1a, b). Using linear mixed-effects modelling of the top five phyla and Shannon's diversity index, we determined three distinct phases of microbiome progression: a developmental phase (months 3–14), a transitional phase (months 15–30), and a stable phase ($\geq$31 months), in which all five phyla and the Shannon diversity index changed significantly during the developmental phase, two phyla (*Proteobacteria* and *Bacteroidetes*) and the Shannon diversity index changed significantly during the transitional phase, and all phyla and the Shannon diversity index were unchanged during the stable phase (Fig. 1c). *Bifidobacterium* dominated during the initial developmental phase, in which 20% of individuals transitioned from cluster 1 to cluster 3 (*Bifidobacterium* was dominant in both clusters). As infants aged, the microbiomes of their stools diversified into clusters 4–8 during months 15–30 (that is, the transitional phase). Microbiome stabilization, in which infants' samples remained in the same cluster at consecutive time points, was observed from month 31 of life. Clusters 8–10 were the most dominant during the stable phase, with these clusters characterized by high alpha diversity and dominance of genera within the Firmicutes phyla. The three microbiome phases and changes in taxa are consistent with other cohorts[15–18] and were supported by the metagenomic sequencing data (Supplementary Note 2 and Extended Data Fig. 2).

We next sought to determine the significant factors associated with the microbiome profiles from 16S rRNA gene sequencing (genus level),

[1]Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA. [2]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK. [3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Barbara Davis Center for Childhood Diabetes, University of Colorado, Aurora, CO, USA. [6]Pacific Northwest Research Institute, Seattle, WA, USA. [7]Institute of Biomedicine, Research Centre for Integrative Physiology and Pharmacology, University of Turku, Turku, Finland. [8]Department of Pediatrics, Turku University Hospital, Turku, Finland. [9]Institute of Diabetes Research, Helmholtz Zentrum München, Munich, Germany. [10]Forschergruppe Diabetes, Technische Universität München, Klinikum Rechts der Isar, Munich, Germany. [11]Forschergruppe Diabetes e.V. at Helmholtz Zentrum München, Munich, Germany. [12]Center for Biotechnology and Genomic Medicine, Medical College of Georgia, Augusta University, Augusta, GA, USA. [13]National Institute of Diabetes & Digestive & Kidney Diseases, Bethesda, MD, USA. [14]Department of Clinical Sciences, Lund University/CRC, Skane University Hospital, Malmö, Sweden. [15]Department of Virology, Faculty of Medicine and Biosciences, University of Tampere, Tampere, Finland. [16]Fimlab Laboratories, Pirkanmaa Hospital District, Tampere, Finland. [17]Health Informatics Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA. [18]These authors contributed equally: Christopher J. Stewart, Nadim J. Ajami. *e-mail: christopher.stewart@ncl.ac.uk; jpetrosi@bcm.edu

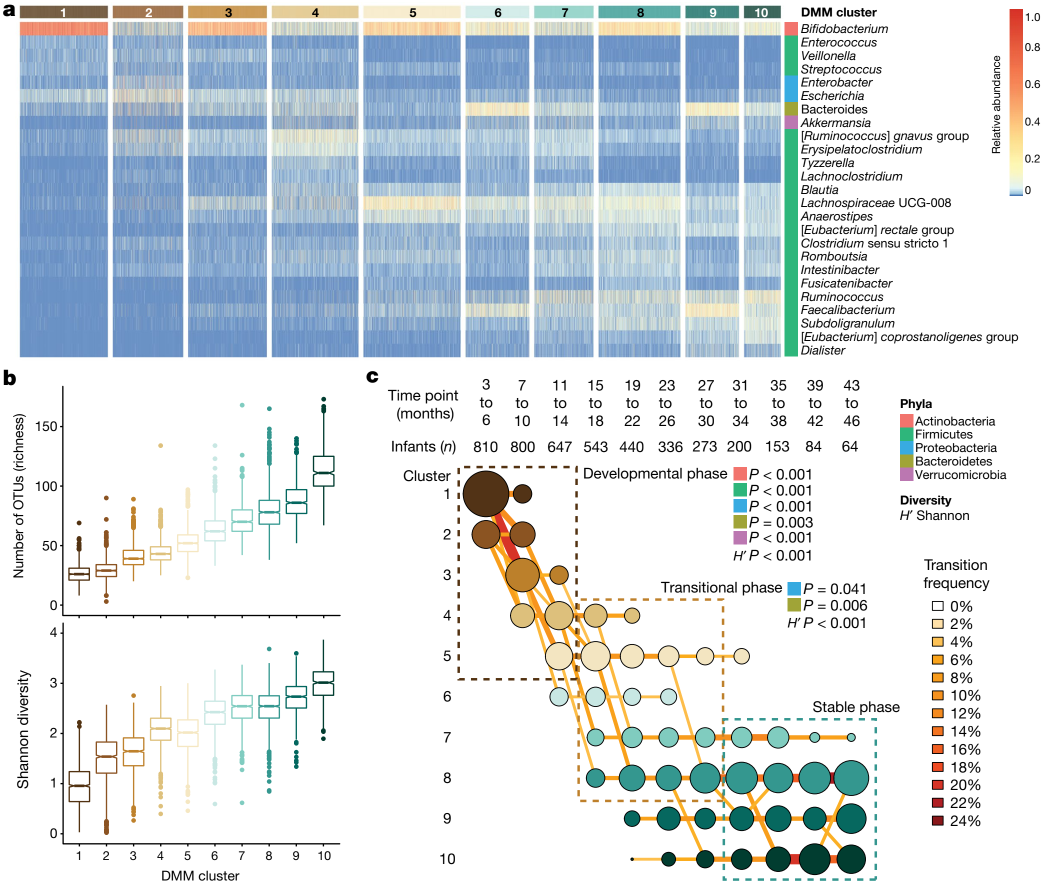


Fig. 1 | DMM clustering of 16S rRNA gene sequencing data ($n = 12{,}005$). The entire dataset formed ten distinct clusters based on lowest Laplace approximation. **a**, Heat map showing the relative abundance of the 25 most dominant bacterial genera per DMM cluster. Taxa names in square brackets are in need of formal taxonomic revision. **b**, Box plots showing the alpha diversity (richness and Shannon's diversity) per each DMM cluster. The centre line denotes the median, the boxes cover the 25th and 75th percentiles, and the whiskers extend to the most extreme data point, which is no more than 1.5 times the length of the box away from the box. Points outside the whiskers represent outlier samples. **c**, Transition model showing the progression of samples through each DMM cluster per each time point, from months 3 to 46 of life. Dashed boxes show the three phases of microbiome progression (developmental, transitional and stable phase). Solid squares next to the labels denote the significant changes in phyla and Shannon's diversity (H') per phase based on multiple linear regression. All phyla and the H' were significant in the developmental phase, two phyla and the H' were significant in the transitional phase, and no phyla or the H' were in the stable phase. Nodes and edges are sized based on the total counts. Nodes are coloured according to DMM cluster number and edges are coloured by the transition frequency. Transitions with less than 4% frequency are not shown. Results are further supported by the metagenomic sequencing data in Extended Data Fig. 2.

metagenomic sequencing taxa (species level), and functional metabolic capacity (Kyoto encyclopedia of genes and genomes (KEGG) modules) (Supplementary Table 1). For statistical analysis, covariates were analysed by stratifying the samples into discrete time points (months 3–6, 7–10, 11–14, 15–18, 19–22, 23–26, 27–30 and 31–40), and only the first sample from each infant was included. Information about the underlying grouping of each covariate is shown in Extended Data Table 1. Several covariates were significantly associated with the genus and species level bacterial community profiles between months 3 and 18 of age, particularly at the first time point of 3 to 6 months (Fig. 2). Conversely, bacterial metabolic potential was associated exclusively with the consumption of breast milk from months 3 to 14 of life (Fig. 2).

Breastfeeding explained the greatest amount of variance from months 3 to 14 of life, after which only 10% of infants received any breast milk (Fig. 2). Breastfeeding had a comparable influence on microbiome development, regardless of whether it was exclusive or together with formula milk and/or solids (Fig. 3a). At the genus level, the receipt of breast milk was most significantly associated with *Bifidobacterium* throughout each time window (Supplementary Table 2). At the species level, breastfeeding was significantly associated with 121 different bacterial species, with higher levels of *B. bifidum*, *B. breve*, *B. dentium*, *Lactobacillus rhamnosus* and *Staphylococcus epidermidis*, and lower levels of *Escherichia coli*, *Tyzzerella nexilis*, *Eggerthella lenta*, *Ruminococcus torques* and *Roseburia intestinalis* in infants that were breastfed (a full list of significant taxa and associated *P* values are shown in Supplementary Table 2). *Bifidobacterium* spp. and *Lactobacillus* spp. exist viably in breast milk and *Staphylococcus* spp. colonize the areolar skin, thus these species can be directly transferred from the mother to infant[19–22]. *B. longum* was not significantly associated with breastfeeding and remained in higher relative abundance compared to other *Bifidobacterium* spp. (Fig. 3b). In the companion manuscript by Vatanen et al.[14], most *B. longum* strains were found to contain genes from the human milk oligosaccharide (HMO) gene cluster, whereas after the cessation of breast milk, most *B. longum* strains no longer carried these genes. This potentially reflects the ability of *B. longum* subsp. *infantis* and subsp. *longum* to use mammalian- and plant-derived oligosaccharides, respectively[23,24]. *B. bifidum* also persisted after the cessation of breastfeeding, and this species is able to switch HMO to mucin degradation[24]. Vatanen et al.[14] show experimentally that *B. breve*, *B. longum* and *B. bifidum*, which make up DMM clusters 1–3 (Extended Data Fig. 2), have distinct profiles of sugar utilization, suggesting that the different nutrient availability between infants can promote the colonization of specific *Bifidobacterium* species.

As the infant ages, the proportion of solid foods in the diet increases (and the amount of breast milk decreases)[21]. In the current study, the Shannon diversity index between infants receiving some breast milk and infants no longer receiving breast milk began to converge over time, probably as a result of a reduced proportion of breast milk in the

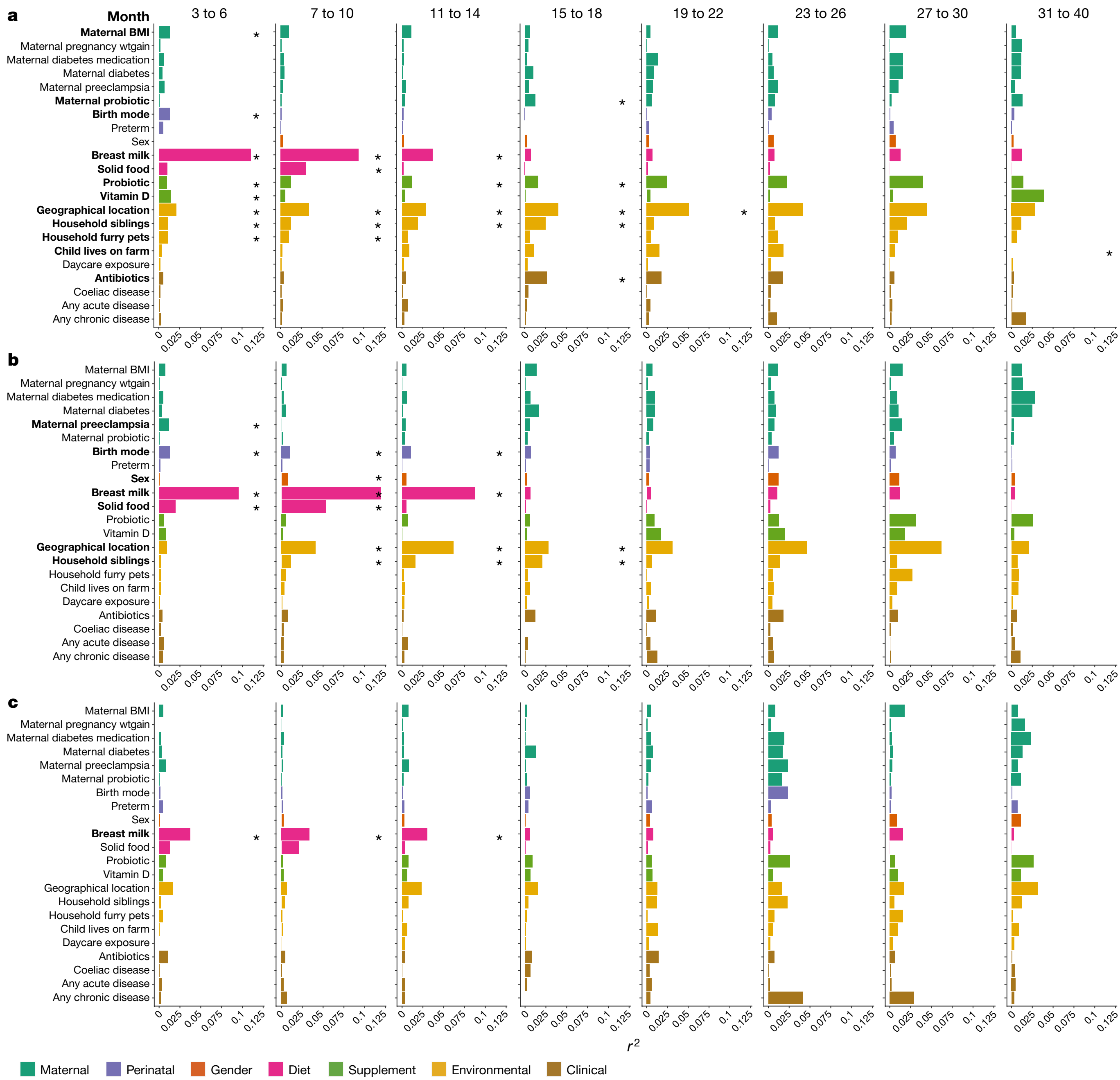


Fig. 2 | Significance and explained variance of 22 microbiome covariates modelled by EnvFit across all data types. Horizontal bars show the amount of variance (r^2) explained by each covariate in the model as determined by EnvFit. The groups within each covariate are detailed in Extended Data Table 1. Covariates are coloured based on overall metadata group. Significant covariates (false discovery rate (FDR) $P < 0.05$) are represented in bold font. Asterisk denotes the significant covariates at each time point. BMI, body mass index; wtgain, weight gain. **a**, Microbiome profiles at the genus level based on 16S rRNA gene sequencing data ($n = 4{,}069$). **b**, Microbiome profiles at the species level based on metagenomic sequencing ($n = 3{,}843$). **c**, Functional metagenomic capacity at the module level based on metagenomic sequencing ($n = 3{,}843$).

diet and therefore less dominance of *Bifidobacterium* (Fig. 3c). Infants receiving some breast milk had significantly lower diversity when compared with infants no longer receiving breast milk across all phases ($P < 0.001$ for all phases), owing to the dominance of *Bifidobacterium* in infants receiving breast milk. To explore microbiome maturation further, we used microbiota age and microbiota-by-age Z-scores (MAZ) as previously described[25], with a model of 20 operational taxonomic units (OTUs) that explained 72% of the variance (compared to 74% when including all OTUs in the model) (Extended Data Fig. 3). Comparably, the microbiota age and MAZ scores were significantly reduced in infants receiving some breast milk in the developmental and transitional phases (both $P < 0.001$ for microbiota age and MAZ scores), but converged in the stable phase (microbiota age $P = 0.331$ and MAZ score $P = 0.196$) (Fig. 3d). After the cessation of breast milk, 110 unique bacterial species (89 from the Firmicutes phylum) were significantly increased from months 3 to 14 of life alone (Supplementary Table 2). The suppression of Firmicutes while in receipt of some breast milk was recently noted[21]. Together, these data support existing reports that the maturation of the gut microbiome is driven by the cessation of breast milk (rather than the introduction of solid foods), hallmarked by increased levels of Firmicutes[17,21,26].

Breastfeeding was the only covariate that was significantly associated with metabolic potential (Fig. 2). Plotting all significant modules (Supplementary Table 1) from the first three time points (months 3–14) showed clear clustering based on the receipt of breast milk, with comparability in the metabolic capacity regardless of the time

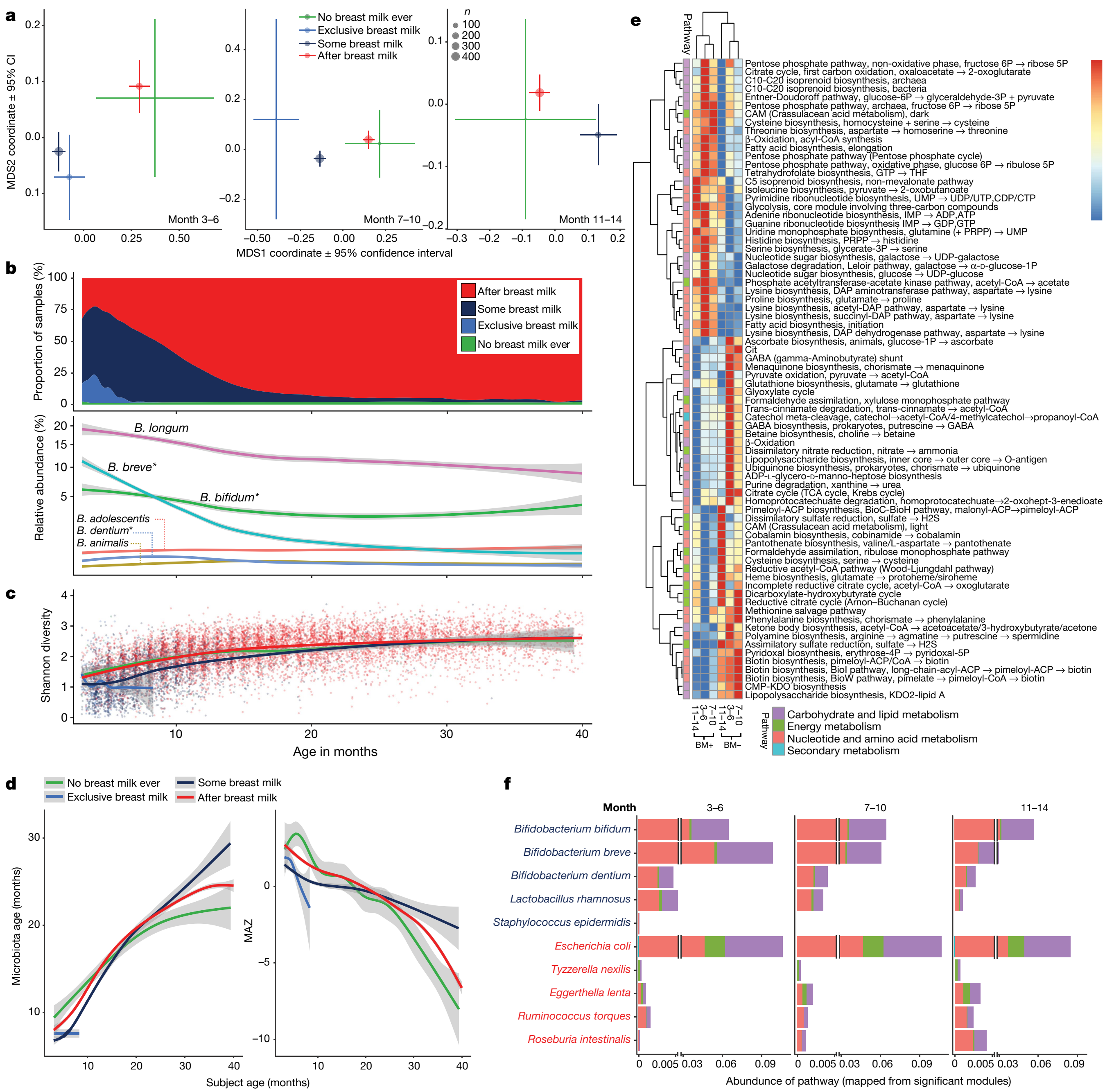


Fig. 3 | Breastfeeding status was the most significant microbiome covariate associated with all datasets throughout the first year of life. Breastfeeding status was significantly associated with microbiome profiles over the first three time points (months 3–14, $n = 2{,}257$; Supplementary Table 1). Curves show locally weighted scatterplot smoothing (LOESS) for the data per category, and shaded areas show permutation-based 95% confidence intervals for the fit. **a**, Non-metric multidimensional scaling (NMDS) ordination plots showing the mean centroid of each breastfeeding status group. Plots include only the first sample obtained from a patient within a given time point; months 3–6, 7–10 and 11–14. Centroid size based on number of samples and the bars represent the ±95% confidence interval. **b**, Plots showing the receipt of breast milk from months 3 to 40 of age compared to the relative abundance of the six most abundant *Bifidobacterium* species over the same period ($n = 11{,}717$). **c**, Longitudinal Shannon diversity index from months 3 to 40 of age ($n = 11{,}717$). **d**, Longitudinal development of the microbiome maturation based on the microbiota age and MAZ score against the age of the infant at sampling ($n = 11{,}717$). **e**, Heat map showing the mean abundance of all significant modules as determined by MaAsLin analysis at each of the first three time points. The corresponding pathway for each module is also presented. BM, breast milk. **f**, Stacked bar plots showing the abundance of each significant module binned at the pathway level. Abundance plotted per bacterial species, with the five most significant species associated with breastfed and non-breastfed infants, respectively.

point (Fig. 3e). Modules most significantly associated with breastfed infants were from the 'carbohydrate and lipid metabolism' pathway and included 'fatty acid biosynthesis' (M00083 and M00082) and 'beta-oxidation, acyl-CoA synthesis' (M00086) (Supplementary Table 2). This is in accordance with previous work that found that genes that relate to the biosynthesis of fatty acids are increased during infancy in breastfed infants[17,27,28]. Conversely, infants not receiving breast milk showed rapid turnover of the metabolic capacity, and the 'dicarboxylate-hydroxybutyrate cycle' (M00374) and 'reductive acetyl-CoA (Wood–Ljungdahl)' (M00377) pathways were increased. Modules relating to vitamins B7 ('nucleotide and amino acid metabolism' pathway; M00573, M00577 and M00123) were also increased in all time points up to 14 months in non-breastfed infants, a function that is associated with the adult microbiome[28].

By mapping reads with genomic coordinates that overlap with known KEGG orthologues to KEGG modules (M), we were able to directly determine from which taxa each gene orthology (and thus module) was derived (see Methods). Each pathway from which each significant module was derived was plotted against the main species discriminating breastfeeding status (Supplementary Table 2). In breastfed infants, *B. breve* accounted for the highest number of significant modules in early life, and was replaced by *B. bifidum* after 6 months of life (Fig. 3f). In non-breastfed infants, *E. coli* primarily accounted for the significant modules between 3 and 14 months of life (Fig. 3f). This provides further evidence that the gut microbiome rapidly matures after the cessation of breast milk, both at the taxonomic and functional levels.

The TEDDY study was powered to detect microbiome associations with the development of IA and T1D based on a specific 1:1 nested case–control study design, from two nested case–control studies (IA or T1D), using risk set sampling[29]. The analytical cohort consisted of a subset with an equal number of samples for each case–control pair. The IA cohort consisted of 632 children and 6,194 stool samples and the T1D cohort consisted of 196 children and 1,540 stool samples, as of 31 May 2012 (Supplementary Table 3). The temporal alpha diversity (both richness and Shannon's diversity), microbiota age and MAZ scores were comparable between cases and matched controls for both the IA and T1D groups (all $P > 0.05$; Extended Data Fig. 4a–h). The relative abundance of the top 50 most abundant genera from 16S rRNA gene sequencing showed only subtle compositional differences, with higher relative abundance of an unclassified Erysipelotrichaceae ($P = 0.019$) in cases of IA (Supplementary Table 3). In the T1D and control cohort, five bacterial genera were associated with T1D onset, with *Parabacteroides* the most significant ($P < 0.001$). Eleven bacterial genera were lower in T1D cases, including four unclassified Ruminococcaceae, *Lactococcus* ($P = 0.020$), *Streptococcus* ($P = 0.032$), and *Akkermansia* ($P = 0.045$) (Supplementary Table 3).

Conditional logistic regression models showed no significant associations between either the numbers of unique states exhibited or the number of transitions between different states per subject for IA (Extended Data Fig. 4i and Supplementary Table 3). The lack of associations was consistent in T1D, with the exception that cases exhibited fewer unique states 6–12 months before the onset of T1D ($P = 0.032$) (Extended Data Fig. 4j and Supplementary Table 3). Notably, the 6–12 months before T1D onset group consisted of the lowest number of samples for any of the time points ($n = 67$ subjects per group), and thus the statistically significant result should be interpreted with caution. Overall, the conditional logistic regression models of community dynamics suggest that microbiome stability was not strongly related to the onset of IA or T1D.

Further analysis of covariates that were significant at several time points and/or consistently significant by 16S rRNA gene sequencing and metagenomics are presented in Supplementary Note 3. In brief, birth mode was significantly associated with microbiome development over the first year of life, with higher levels of *Bacteroides* spp. in infants that were delivered vaginally (Extended Data Fig. 5). This was generally consistent across the different breast milk exposure groups and geographical locations (Extended Data Fig. 6). Differences between geographical locations occurred from 3 to 22 months of life (Supplementary Table 1), although the core microbiome was consistent (Supplementary Table 4), and diversity, microbiota age and MAZ scores had comparable trajectories across each location (Extended Data Fig. 7a–c). Household exposures (for example, living with siblings and with furry pets) were also associated with differences in the microbiome profiles in early life, in which infants living with siblings and/or with furry pets showed accelerated rates of maturation of the microbiome (Extended Data Fig. 7d–i).

The TEDDY population offers a robust analysis of gut microbiome development of 903 infants from months 3 to 46 of age, with regular sampling (more than 12,000 stool samples), extensive metadata, and the use of both amplicon and metagenomic sequencing. We showed that the first year of life is a key phase for the development of the microbiome, with the receipt of breast milk being the main factor that influences microbiome development over this period. Birth mode, geographical location, household siblings and furry pets were also associated with the microbiome over this period. We considered the first year of life as developmental, the second year of life as transitional, and from year three of life the microbiome stabilized. These precise ages may shift when investigators include samples before month 3 or beyond month 46 of life.

The current cohort is largely white, non-Hispanic and is drawn from a population of infants at high genetic risk for T1D, some of whom developed autoimmunity or diabetes. Temporal alpha diversity and community dynamics were comparable between cases and controls, which is in contrast to findings reported in other cohorts and may reflect the increased number of subjects and samples in the TEDDY cohort[11,12]. We found subtle changes in the relative abundance of bacterial genera between cases (IA and/or T1D) and matched controls. T1D cases showed higher levels of *Streptococcus* sp. and *Lactococcus* sp., which is consistent with the findings of Vatanen et al.[14] in the companion paper. In accordance with previous work, the abundance of *Akkermansia* was also higher in controls in the current study, which may be indicative of enhanced gut integrity[10].

The overall microbiome development and significant covariates are in concordance with previous reports in westernized populations, although caution should be exercised when extrapolating the findings from the TEDDY cohort of children with risk factors of developing T1D to the wider population. Nevertheless, the significant covariates reported in the current study have been independently linked to the risk of later life diseases such as obesity, asthma and allergy[1–8]. The current study provides several testable hypotheses of microbiome development in infancy, and it remains important to determine the potential mechanism of altered early life microbiome and the subsequent effect on immune development and functioning. With a more comprehensive understanding of the crucial early life phases and their effect on health and disease, lifestyles and therapeutics can be tailored to support optimal microbial–immune homeostasis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0617-x.

Received: 16 November 2017; Accepted: 30 August 2018;
Published online 24 October 2018.

1. Yuan, C. et al. Association between cesarean birth and risk of obesity in offspring in childhood, adolescence, and early adulthood. *JAMA Pediatr.* **170**, e162385 (2016).
2. Sevelsted, A., Stokholm, J., Bønnelykke, K. & Bisgaard, H. Cesarean section and chronic immune disorders. *Pediatrics* **135**, e92–e98 (2015).
3. Mayer-Davis, E. J. et al. Breast-feeding and risk for childhood obesity: does maternal diabetes or obesity status matter? *Diabetes Care* **29**, 2231–2237 (2006).
4. Klement, E., Cohen, R. V., Boxman, J., Joseph, A. & Reif, S. Breastfeeding and risk of inflammatory bowel disease: a systematic review with meta-analysis. *Am. J. Clin. Nutr.* **80**, 1342–1352 (2004).
5. Koplin, J. J. et al. Environmental and demographic risk factors for egg allergy in a population-based study of infants. *Allergy* **67**, 1415–1422 (2012).
6. Benn, C. S., Melbye, M., Wohlfahrt, J., Björkstén, B. & Aaby, P. Cohort study of sibling effect, infectious diseases, and risk of atopic dermatitis during first 18 months of life. *Br. Med. J.* **328**, 1223 (2004).
7. Hesselmar, B., Åberg, N., Åberg, B., Eriksson, B. & Björkstén, B. Does early exposure to cat or dog protect against later allergy development? *Clin. Exp. Allergy* **29**, 611–617 (1999).
8. Virtanen, S. M. et al. Microbial exposure in infancy and subsequent appearance of type 1 diabetes mellitus-associated autoantibodies: a cohort study. *JAMA Pediatr.* **168**, 755–763 (2014).
9. Aagaard, K., Stewart, C. J. & Chu, D. Una destinatio, viae diversae: does exposure to the vaginal microbiota confer health benefits to the infant, and does lack of exposure confer disease risk? *EMBO Rep.* **17**, 1679–1684 (2016).
10. Brown, C. T. et al. Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes. *PLoS ONE* **6**, e25792 (2011).
11. Giongo, A. et al. Toward defining the autoimmune microbiome for type 1 diabetes. *ISME J.* **5**, 82–91 (2011).
12. Kostic, A. D. et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).

13. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) study: study design. *Pediatr. Diabetes* **8**, 286–298 (2007).
14. Vatanen, T. et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* https://doi.org/10.1038/s41586-018-0620-2 (2018).
15. Yassour, M. et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med*. **8**, 343ra81 (2016).
16. Bokulich, N. A. et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med*. **8**, 343ra82 (2016).
17. Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 690–703 (2015).
18. Penders, J. et al. Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* **118**, 511–521 (2006).
19. Martín, R. et al. Isolation of bifidobacteria from breast milk and assessment of the bifidobacterial population by PCR-denaturing gradient gel electrophoresis and quantitative real-time PCR. *Appl. Environ. Microbiol*. **75**, 965–969 (2009).
20. Hunt, K. M. et al. Characterization of the diversity and temporal stability of bacterial communities in human milk. *PLoS ONE* **6**, e21313 (2011).
21. Pannaraj, P. S. et al. association between breast milk bacterial communities and establishment and development of the infant gut microbiome. *JAMA Pediatr*. **171**, 647–654 (2017).
22. Soeorg, H. et al. The role of breast milk in the colonization of neonatal gut and skin with coagulase-negative staphylococci. *Pediatr. Res*. **82**, 759–767 (2017).
23. Underwood, M. A., German, J. B., Lebrilla, C. B. & Mills, D. A. *Bifidobacterium longum* subspecies *infantis*: champion colonizer of the infant gut. *Pediatr. Res*. **77**, 229–235 (2015).
24. O'Callaghan, A. & van Sinderen, D. Bifidobacteria and their role as members of the human gut microbiota. *Front. Microbiol*. **7**, 925 (2016).
25. Subramanian, S. et al. Persistent gut microbiota immaturity in malnourished Bangladeshi children. *Nature* **510**, 417–421 (2014).
26. Bergström, A. et al. Establishment of intestinal microbiota during early life: a longitudinal, explorative study of a large cohort of Danish infants. *Appl. Environ. Microbiol*. **80**, 2889–2900 (2014).
27. Koenig, J. E. et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl Acad. Sci. USA* **108**, 4578–4585 (2011).
28. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
29. Lee, H. S. et al. Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Diabetes Metab. Res. Rev*. **30**, 424–434 (2014).

Acknowledgements We acknowledge the following members of the CMMR for their support in samples processing: T. Ayvaz, T. Bauch, L. Kusic, L. Railey, R. Berry, A. Tamegnon, E. Zavala, H. Moreno and N. Truong. In addition, we acknowledge the contribution of the Human Genome Sequencing Center at Baylor College of Medicine for their support in the data generation aspects of this work. We apologize to authors of existing work that we could not cite because of space constraints. This research was performed on behalf of the TEDDY Study Group, which is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, UC4 DK100238, UC4 DK106955, UC4 DK112243, UC4 DK117483, and contract no. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. This work supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida (UL1 TR000064) and the University of Colorado (UL1 TR001082). R.J.X. was supported by funding from JDRF (2-SRA-2016-247-S-B and 2-SRA-2018-548-S-B).

Reviewer information *Nature* thanks K. Aagaard, C. Lozupone and L. Wen for their contribution to the peer review of this work.

Author contributions C.J.S., N.J.A., R.E.L., T.V., C.H., R.J.X., M.R., W.H., J.T., A.-G.Z., J.-X.S., B.A., A.L., H.H., K.V., J.P.K. and J.F.P. designed the study; M.R., W.H., J.T., A.-G.Z., J.X.S., B.A., A.L., H.H., K.V. and J.P.K. participated in patient recruitment and diagnosis, sample collection, generation of the metadata; C.J.S., N.J.A., M.C.W., M.C.R., H.D., G.A.M., D.M. and R.A.G. generated and processed the raw sequencing data; C.J.S., N.J.A., J.L.O., D.S.H. and D.P.S. performed the data analysis, data interpretation, and figure generation; C.J.S., N.J.A., J.L.O., D.S.H. and J.F.P. wrote the paper; and all authors contributed to critical revisions and approved the final manuscript. Members of the TEDDY Study Group are listed in the Supplementary Information.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0617-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0617-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.J.S. or J.F.P.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Study population. The TEDDY Study is composed of six clinical research centres: three in the United States (Colorado, Georgia/Florida and Washington), and three in Europe (Finland, Germany and Sweden). Children enrolled are followed prospectively from three months to 15 years with study visits every three months until age 4 years and every three or six months thereafter depending on autoantibody positivity. Stool samples and associated metadata were collected as of 31 May 2012. Stool samples were collected monthly from 3 to 48 months of life, then every three months until the age of 10 years, and then biannually thereafter, into the three plastic stool containers provided by the clinical centre. Children who were antibody negative after 4 years of age were encouraged to submit four times a year even though after 4 years their visits schedule switched to biannual. Parents sent the stool containers at either ambient or +4 °C temperature with guaranteed delivery within 24 h in the appropriate shipping box to the NIDDK repository if living in the United States or their affiliated clinical centre if living in Europe. The European clinical centres stored the stool samples and sent monthly bulk shipments of frozen stool to the NIDDK repository. The population (both cases and controls) is based on children at high risk for T1D based on their HLA genotype with 10% based on family history in addition to HLA. Detailed study design and methods have been previously published[13,29,30]. Matching factors for case and control children were geographical location, sex and family history of T1D.

Metadata were collected using validated questionnaires that have been either published or extensively scrutinized by experts. Information about mothers, pregnancy and birth was collected during the three month clinic visit by questionnaire and included the mode of birth (vaginal birth versus Caesarean section), the infant's 5-min Apgar score, pregnancy complications, information about maternal diabetes (T1D, type 2 diabetes (T2D) or gestational diabetes), gestational age, and maternal medication use (insulin, metformin, glyburide, antihypertensives) during pregnancy. TEDDY provides many tools, such as 'The TEDDY book', to the parents to assist in real-time collection of all events in their child's life to ensure bias and error are minimized. At each visit the study personnel will go over the TEDDY book with the primary caretaker and extract pertinent information using standardized study forms. Data are extracted by trained staff members during scheduled visits every three months starting at 3 months of age and entered directly via stand forms (web forms or teleforms), which are transmitted electronically. Front-end constraints are used in the web application to prevent the entry of invalid data and The TEDDY Error Reporting and Verification System (ERVS) consists of a set of programs that conduct automated quality control on the data, report and resolve errors, an integrated database for storing error data, and a set of programs that generate reports for monitoring data cleaning efforts. The details of the system have been published[31]. Given the prospective nature of the TEDDY design, information and recall bias are greatly minimized. Because the children do not have event outcome at time of enrolment and are followed, there is no reason for any systematic differences between groups of the study participants in the accuracy of the information collected.

The TEDDY study was approved by local US Institutional Review Boards and European Ethics Committee Boards in Colorado's Colorado Multiple Institutional Review Board, Georgia's Medical College of Georgia Human Assurance Committee (2004–2010), Georgia Health Sciences University Human Assurance Committee (2011–2012), Georgia Regents University Institutional Review Board (2013–2015), Augusta University Institutional Review Board (2015–present), Florida's University of Florida Health Center Institutional Review Board, Washington state's Washington State Institutional Review Board (2004–2012) and Western Institutional Review Board (2013–present), Finland's Ethics Committee of the Hospital District of Southwest Finland, Germany's Bayerischen Landesärztekammer (Bavarian Medical Association) Ethics Committee, Sweden's Regional Ethics Board in Lund, Section 2 (2004–2012) and Lund University Committee for Continuing Ethical Review (2013–present). All parents or guardians provided written informed consent before participation in genetic screening and enrolment. The study was performed in compliance with all relevant ethical regulations.

A priori power calculations using discrete Cox's proportional hazards regression[32] for the matched IA case–control study estimated 80% power, $\alpha = 0.01$, two-sided test to detect an odds ratio > 3 for an exposure with 5% prevalence to an odds ratio > 1.8 for an exposure with 20% prevalence. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

16S rRNA gene sequencing. 16S rRNA gene sequencing methods were adapted from the methods developed by the NIH-Human Microbiome Project and the Earth Microbiome Project[33–35]. Bacterial DNA was extracted using the PowerMag Microbiome DNA isolation kit following the manufacturer's instructions. The V4 region of the 16S rRNA gene was amplified by PCR and sequenced on the MiSeq platform (Illumina) using the 2 × 250 bp paired-end read protocol. The read pairs were demultiplexed and reads were merged using USEARCH v7.0.1090[36]. Merging allowed zero mismatches and a minimum overlap of 50 bases, and merged reads were trimmed at the first base with a $q \leq 5$. A quality filter was applied to the resulting merged reads and those containing above 0.5% expected errors were discarded. Sequences were stepwise clustered into OTUs at a similarity cut-off value of 97% using the UPARSE algorithm[37]. Chimeras were removed using USEARCH v7.0.1090 and UCHIME v4.2. To determine taxonomies, OTUs were mapped to a version of the SILVA Database[38] containing only the 16S V4 region using USEARCH v7.0.1090. Abundances were recovered by mapping the merged reads to the UPARSE OTUs. A custom script constructed a rarefied OTU table from the output files generated in the previous two steps for downstream analyses of taxonomic relative abundance, alpha diversity, and beta diversity (including UniFrac)[39]. A total of 114,313,601 reads (median 8,442 reads per sample) were obtained from 16S rRNA gene sequencing and each sample was rarefied to 3,000 reads. Stringent merging parameters account for the relatively low number of OTUs, with the number of species by metagenomics around fourfold higher than the number of OTUs by 16S rRNA gene sequencing.

Metagenomic shotgun sequencing. Individual libraries constructed from each sample were pooled and loaded onto the HiSeq 2000 platform (Illumina) and sequenced using the 2 × 100 bp paired-end read protocol. The process of quality filtering, trimming, and demultiplexing was carried out by in-house pipeline developed by assembling publicly available tools such as Casava v1.8.2 (Illumina) for the generation of fastqs, Trim Galore v0.2.8 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and cutadapt v1.9dev2 for adaptor and quality trimming, and PRINSEQ v0.20.5[40] for sample dereplication and low complexity filtering. In addition, Bowtie2 v2.2.3[41] was used to map reads to a database containing complete genomes and assemblies for bacteria, viruses, human, and vectors in the NCBI whole-genome sequencing (WGS) archive (as of March 2015). Reads in which the highest identity matches were not bacterial were removed from subsequent analysis. The edit distance (Levenshtein distance) was used to determine the score of the alignments to the reference genomes[42]. For bacterial reads, the highest scoring match (greater than 90%) was chosen per read considering only the top 25 highest scoring alignments. In the event of multiple identical top scoring hits, the lowest common ancestor was determined.

Reads in which the genomic coordinates overlap with known KEGG orthologues[43,44] were tabulated, and KEGG modules were calculated step-wise and determined to be complete if 65% of the reaction steps were present per detected species and for the metagenome. Pathways were constructed for each taxa and metagenome by calculating the minimum set through MinPath[45] resulting from the gene orthologues present. A total of 19,967,936,136 reads (median 1,606,240 reads per sample) were obtained from metagenomic sequencing and for subsequent analysis each sample was rarefied to 100,000 reads.

Statistical analysis. The analysis was conducted in two parts: (1) characterize the longitudinal maturation of the microbiome and (2) determine the significant covariates that influence microbiome development. For both parts of analysis, alpha diversity (richness and Shannon diversity) was calculated at the OTU-level for 16S rRNA gene sequencing and species-level for metagenomics data. Alpha diversity and taxonomic abundance were modelled using LOESS regression, and implemented and plotted with 95% confidence intervals in R (http://www.R-project.org) using the ggplot package[46].

DMM clustering. The first part of the analysis determined the key phases of microbiome progression, which included the use of DMM. DMM bins samples on the basis of microbial community structure[47]. The appropriate number of clusters was determined based on the lowest Laplace approximation score. For this specific analysis, samples up to month 46 of life were included, whereas all other analyses included samples up to month 40 of life. Including the additional samples here allowed for more accurate determination of the microbiome phases.

The second part of the analysis sought to determine the significant covariates in shaping the microbiome profiles at discrete time points and further ascertain the significantly altered taxa based on samples up to month 40 of life. The framework for the statistical analysis considered the longitudinal nature of the dataset and accounted for the dynamic nature of the covariates. Owing to the potential that some covariates might influence the microbiome before the start date (for example, underlying indication for an antibiotic prescription) and some covariates will alter the microbiome for an unknown time frame (for example, microbiome disrupted by antibiotics may continue to be altered months after treatment), covariates were classified as 'before', 'during', or 'after'. In the case a covariate was negative for an infant, all samples would be classified as 'never'. In instances in which several onsets of a covariate were possible (for example, multiple antibiotic start and end time points), after the first onset the covariate was classified as 'after' for the remaining samples, unless another event occurred, in which case 'during' would be applied where appropriate according to the start and stop dates. Analysis was performed at specific time windows, including samples collected between months 3–6, 7–10, 11–14, 15–18, 19–22, 23–26, 27–30 and 31–40. Only the first sample collected from a given child was included in each time window to account for repeated measures.

EnvFit analysis to determine significant covariates. The effect size and significance of each covariate were determined using the 'envfit' function in 'vegan' (https://cran.r-project.org/web/packages/vegan/index.html) comparing the difference in the centroids of each group relative to the total variation. Ordination was performed using NMDS based on Bray–Curtis dissimilarity. The significance value was determined based on 10,000 permutations. All *P* values derived from envfit were adjusted for multiple comparisons using FDR adjustment (Benjamini–Hochberg procedure)[48]. In total, 22 covariates with known associations to gut microbiome development in neonates, infants, and children were included in the envfit analysis and the grouping used for within each variable is presented in Extended Data Table 1. Specifically, we tested maternal factors including diabetes (gestational, T1D, T2D or none)[49], diabetes medication (insulin, metformin, glyburide, antihypertensives)[50], BMI[51,52], gestational weight gain category (excess or non-excess)[53], preeclampsia[52], maternal probiotic consumption[54], as well as offspring factors such as prematurity[18,55], birth mode[15–17,56], gender[56], receipt of breast milk and/or formula[17,53,57–59], introduction of solid foods[60,61], geographical location[57], probiotics[62], vitamin D supplementation[63], antibiotics[18], household siblings[56,64], household furry pets[64,65], living on a farm with animals[66,67], day-care exposure[68], coeliac disease[69], acute disease, and chronic disease[69].

MaAsLin analysis to determine significant taxa associated with each covariate. MaAsLin was used for adjustment of covariates when determining the significance of taxa (genus level for 16S rRNA gene sequencing and species level for metagenomic sequencing) contributing to a specific variable, while accounting for potentially confounding covariates[70]. In brief, this multivariate linear modelling system for microbial data selects from among a set of (potentially high-dimensional) covariates to associate with microbial taxon or pathway abundances. Mixed-effects linear models using a variance-stabilizing arcsin square root transform on relative abundances are then used to determine the significance of putative associations from among this reduced set. Nominal *P* values across all associations are then adjusted using the Benjamini–Hochberg FDR method. Here, microbial features with corrected $q < 0.25$ were reported. All 22 covariates tested in the envfit were included in the adjustment regardless of significance by envfit. Subject age was also included to adjust for potential age driven changes in taxa within each three-month time window and IA and T1D outcome were included to adjust for the nested case control nature of the cohort. The default MaAsLin parameters were applied (maximum percentage of samples NA in metadata 10%, minimum percentage relative abundance 0.01%, $P < 0.05$, $q < 0.25$). All P values were adjusted for multiple comparisons using FDR[48].

Microbiota maturation modelling and linear mixed-effects analysis. The random forest regression model[71] was performed as previously described[25], using the 'randomForest' R package[72]. In brief, the model was trained on 150 randomly selected full term (>37 weeks gestation), vaginally delivered, breastfed infants who had a minimum of 10 samples included in the final dataset. The model was built using the default parameters: growing 10,000 trees and $n/3$ OTUs randomly sampled at each split, in which n represents the number of OTUs. The model was further refined by applying 'rfcv' with tenfold cross-validation resulting in the inclusion of 20 OTUs to train the final model based on percentage increase in mean-squared error. These 20 OTUs explained 72% of the total variance of the model (compared to 75% with all OTUs included). The age of the subject predicted by this model was termed microbiota age and was further used to determine MAZ scores using the formulae described preiously[25]. Significant differences in alpha diversity, microbiota age, and MAZ scores were calculated using linear mixed-effects models in R, with the 'lmer' command within the 'lme4' package[73]. We included random slopes and intercept for individual children, and evaluated delivery mode, age, *Bacteroides* positive or negative, predominant diet, geographical location, presence of siblings, and presence of household pets as fixed effects. To perform these piecewise longitudinal models, we divided samples into the three developmental phases (<14 months, >15–<30 months, and >31 months). Owing to the relatively low number of samples in the exclusive and never breastfed groups, the analysis of breast milk status was conducted based on 'some breast milk' or 'after breast milk', with these groups found to cluster with exclusive and never breastfed, respectively.

Determination of the datasets for IA and T1D nested case–control stability analyses. The development of persistent confirmed IA was assessed every three months. Persistent autoimmunity was defined by the presence of confirmed islet autoantibody on two or more consecutive visits. The date of persistent autoimmunity was defined as the draw date of the first sample of the two consecutive samples that deemed the child persistent confirmed positive for a specific autoantibody (or any autoantibody). T1D was defined according to American Diabetes Association criteria for diagnosis[74]. A dataset with equal numbers of cases and control samples was created to preform conditional logistic regression of summary metric variables (that is, counting for each person the number of unique clusters exhibited and the number of temporal transitions between different clusters). On average, cases tended to have more samples than controls, and therefore had more transitions and observed states, which resulted in spurious associations between our metrics and disease outcome. For this purpose, we created a dataset in which case and control samples were matched to the paired case based on the nearest sample by day of life (unmatched sample or sample outside of ±20% were omitted from analyses). This resulted in an analytical cohort of 316 IA cases and 316 paired controls ($n = 3{,}097$ stool samples in each group) and 98 T1D cases and 98 paired controls ($n = 1{,}270$ stool samples in each group). For consistency, we used these datasets for all matched case–control analyses. The IA and T1D analysis was based on 16S rRNA gene sequencing data only and analysis of the metagenomic sequencing data (that is, species level taxonomic profiling and functional capacity) are presented in the companion paper[14].

Taxonomic and metabolic profiling relative to IA onset in the matched case–control dataset. 16S rRNA gene sequencing data was used to determine differences between alpha diversity (number of OTUs (richness) and Shannon's diversity index), microbiota age, and MAZ scores. Significant differences in alpha diversity, microbiota age, and MAZ scores were calculated using linear mixed-effects models in R, with the 'lmer' command within the 'lme4' package[73]. To perform these piecewise longitudinal models, we divided samples into the three developmental phases (<14 months, >15–<30 months, and >31 months). Conditional logistic regression of matched case–control pairs was performed on the top 50 most dominant bacterial genera from samples prior to disease diagnosis. Odds ratios were calculated with 95% confidence intervals, adjusted for potential confounding variables, including age at sample collection, HLA genotype, mode of delivery, and duration of breastfeeding. Abundance information for genera was entered into the model as $\log_2$-transformed read counts. A value of 0.01 was added to avoid 0 s. The Benjamini–Hochberg procedure was applied to correct for multiple comparisons[48] and corrected $P < 0.05$ was considered significant.

Assessment of microbiome instability based on DMM clusters between IA or T1D cases and controls. For each subject, the total number of clusters exhibited throughout sampling per infant and the number of transitions between different clusters from one sample to the next were calculated to provide summary measures of microbiome stability over time. These summary metrics were then used in conditional logistic regression to assess the relationship of microbiome stability with IA and T1D. Odds ratios were calculated with 95% confidence intervals, adjusted for potential confounding variables, including HLA genotype, mode of delivery, duration of breastfeeding, number of antibiotic courses, and number of infectious episodes.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

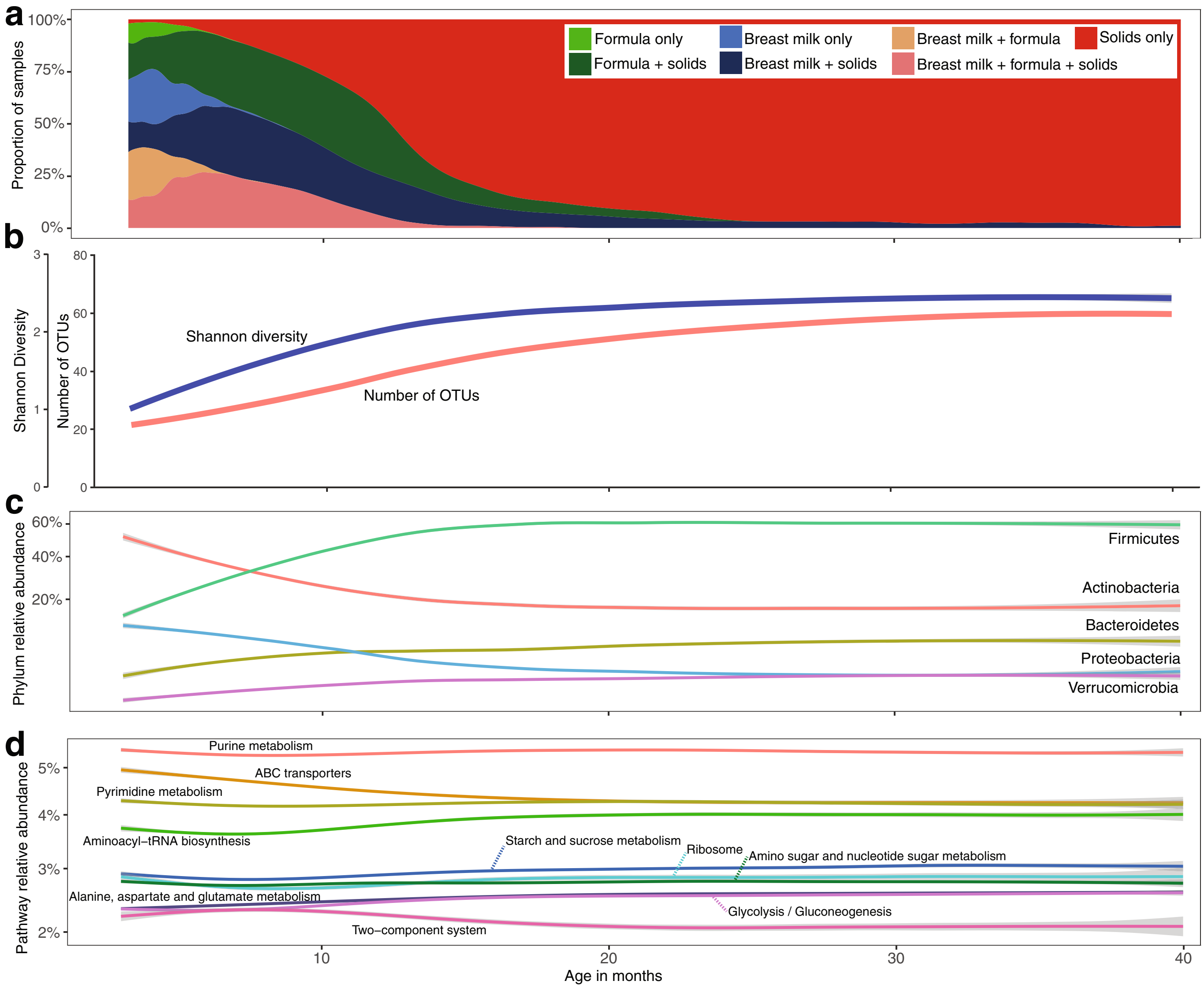
Code availability. Code for the transition model showing the progression of samples through each DMM cluster, which are presented in Fig. 1 and Extended Data Fig. 2, has been made publicly available at https://github.com/StewartLab/Stewart_TEDDY_Microbiome_Analysis. Other analysis software including quality control, taxonomic, and functional profilers is publicly available and referenced as appropriate.

Data availability

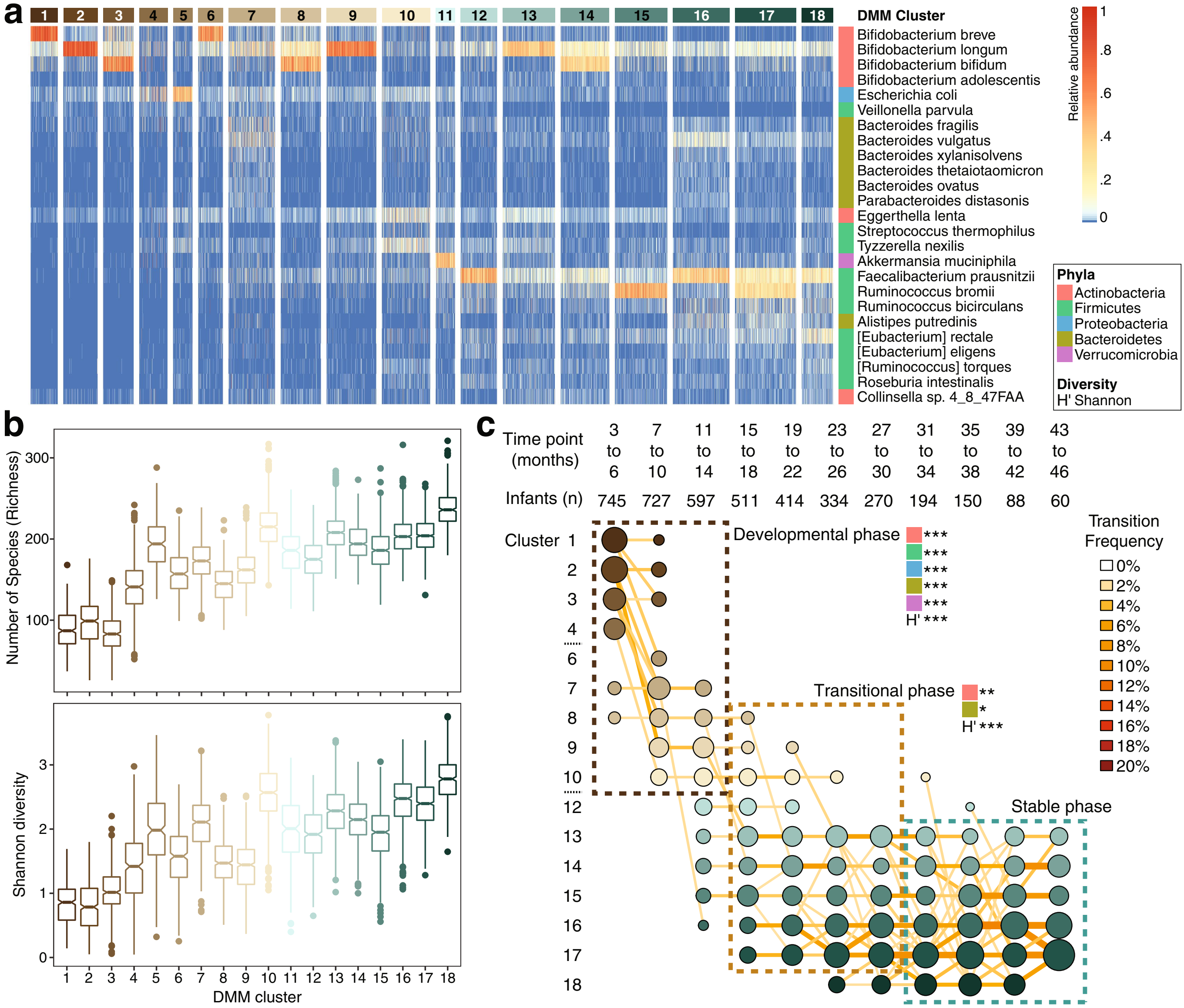
TEDDY microbiome 16S rRNA gene sequencing and metagenomic sequencing data that support the findings of this study have been deposited in the NCBI database of Genotypes and Phenotypes (dbGaP) with the primary accession code phs001443.v1.p1, in accordance with the dbGaP controlled-access authorization process. Clinical metadata analysed during the current study will be made available in the NIDDK Central Repository at https://www.niddkrepository.org/studies/teddy.

30. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) Study. *Ann. NY Acad. Sci.* **1150**, 1–13 (2008).
31. Vehik, K. et al. Methods, quality control and specimen management in an international multi-center investigation of type 1 diabetes: TEDDY. *Diabetes Metab. Res. Rev.* **29**, 557–567 (2013).
32. Lachin, J. M. Sample size evaluation for a multiply matched case–control study using the score test from a conditional logistic (discrete Cox PH) regression model. *Statist. Med.* **27**, 2509–2534 (2012).
33. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
34. The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
35. Caporaso, J. G. et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
36. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
37. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10**, 996–998 (2013).
38. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
39. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).
40. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).

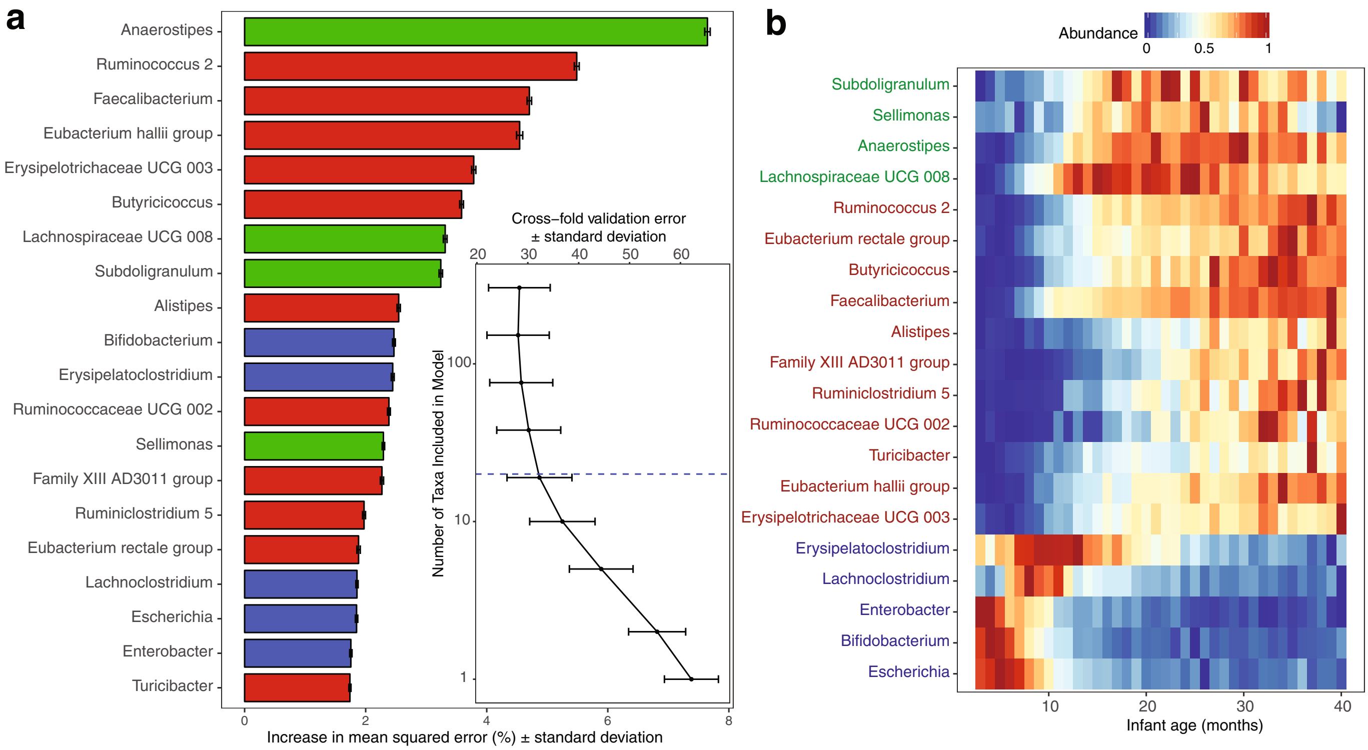
41. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
42. Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl*. **10**, 707–710 (1966).
43. Ogata, H. et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. **27**, 29–34 (1999).
44. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. **28**, 27–30 (2000).
45. Ye, Y. & Doak, T. G. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLOS Comput. Biol*. **5**, e1000465 (2009).
46. Wickham, H. *ggplot2 Elegant Graphics for Data Analysis* (Springer, New York, 2009).
47. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS ONE* **7**, e30126 (2012).
48. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
49. Hu, J. et al. Diversified microbiota of meconium is affected by maternal diabetes status. *PLoS ONE* **8**, e78257 (2013).
50. Wu, H. et al. Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nat. Med*. **23**, 850–858 (2017).
51. Mueller, N. T. et al. Birth mode-dependent association between pre-pregnancy maternal weight status and the neonatal intestinal microbiome. *Sci. Rep*. **6**, 23133 (2016).
52. Gohir, W., Ratcliffe, E. M. & Sloboda, D. M. Of the bugs that shape us: maternal obesity, the gut microbiome, and long-term disease risk. *Pediatr. Res*. **77**, 196–204 (2015).
53. Chu, D. M. et al. Maturation of the infant microbiome community structure and function across multiple body sites and in relation to mode of delivery. *Nat. Med*. **23**, 314–326 (2017).
54. Gueimonde, M. et al. Effect of maternal consumption of lactobacillus GG on transfer and establishment of fecal bifidobacterial microbiota in neonates. *J. Pediatr. Gastroenterol. Nutr*. **42**, 166–170 (2006).
55. Stewart, C. J. et al. Preterm gut microbiota and metabolome following discharge from intensive care. *Sci. Rep*. **5**, 17141 (2015).
56. Martin, R. et al. Early-life events, including mode of delivery and type of feeding, siblings and gender, shape the developing gut microbiota. *PLoS ONE* **11**, e0158498 (2016).
57. Fallani, M. et al. Intestinal microbiota of 6-week-old infants across Europe: geographic influence beyond delivery mode, breast-feeding, and antibiotics. *J. Pediatr. Gastroenterol. Nutr*. **51**, 77–84 (2010).
58. Azad, M. B. et al. Impact of maternal intrapartum antibiotics, method of birth and breastfeeding on gut microbiota during the first year of life: a prospective cohort study. *J. Obstet. Gynaecol*. **123**, 983–993 (2016).
59. La Rosa, P. S. et al. Patterned progression of bacterial populations in the premature infant gut. *Proc. Natl Acad. Sci. USA* **111**, 12522–12527 (2014).
60. Laursen, M. F. et al. Infant gut microbiota development is driven by transition to family foods independent of maternal obesity. *MSphere* **1**, e00069-e15 (2016).
61. Schloss, P. D., Iverson, K. D., Petrosino, J. F. & Schloss, S. J. The dynamics of a family's gut microbiota reveal variations on a theme. *Microbiome* **2**, 25 (2014).
62. Abdulkadir, B. et al. Routine use of probiotics in preterm infants: longitudinal impact on the microbiome and metabolome. *Neonatology* **109**, 239–247 (2016).
63. Sordillo, J. E. et al. Factors influencing the infant gut microbiome at age 3–6 months: findings from the ethnically diverse Vitamin D Antenatal Asthma Reduction Trial (VDAART). *J. Allergy Clin. Immunol*. **139**, 482–491 (2017).
64. Song, S. J. et al. Cohabiting family members share microbiota with one another and with their dogs. *eLife* **2**, e00458 (2013).
65. Tun, H. M. et al. Exposure to household furry pets influences the gut microbiota of infant at 3–4 months following various birth scenarios. *Microbiome* **5**, 40 (2017).
66. Normand, A.-C. et al. Airborne cultivable microflora and microbial transfer in farm buildings and rural dwellings. *Occup. Environ. Med*. **68**, 849–855 (2011).
67. Ege, M. J. et al. Exposure to environmental microorganisms and childhood asthma. *N. Engl. J. Med*. **364**, 701–709 (2011).
68. Thompson, A. L., Monteagudo-Mera, A., Cadenas, M. B., Lampl, M. L. & Azcarate-Peril, M. A. Milk- and solid-feeding practices and daycare attendance are associated with differences in bacterial diversity, predominant communities, and metabolic and immune function of the infant gut microbiome. *Front. Cell. Infect. Microbiol*. **5**, 3 (2015).
69. Aron-Wisnewsky, J. & Clément, K. The gut microbiome, diet, and links to cardiometabolic and chronic disorders. *Nat. Rev. Nephrol*. **12**, 169–181 (2016).
70. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*. **13**, R79 (2012).
71. Breiman, L. Random forests. *Mach. Learn*. **45**, 5–32 (2001).
72. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).
73. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
74. American Diabetes Association. 2. Classification and diagnosis of diabetes. *Diabetes Care* **38**, S8–S16 (2015).



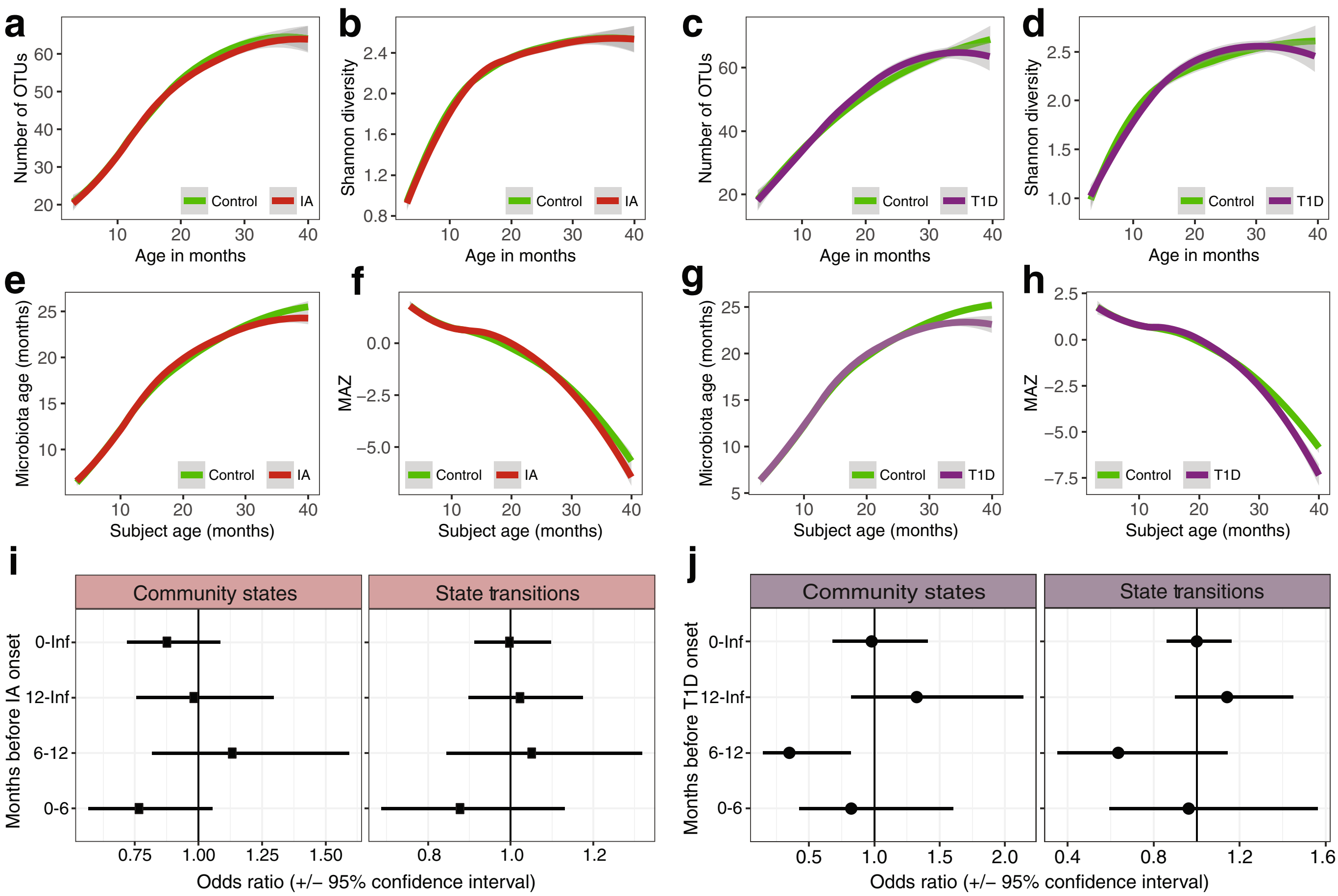
Extended Data Fig. 1 | Characterization of the gut microbiome over the first 40 months of life ($n = 11,717$). a–d, 16S rRNA gene sequencing (**a**–**c**) and metagenomic sequencing (**d**) analysis. Curves show LOESS fit for the data per category, and shaded areas show permutation-based 95% confidence intervals for the fit. **a**, Summary of overall dietary status. **b**, The mean alpha diversity (richness and Shannon diversity) per child increased rapidly from 3 to 20 months of life. **c**, The mean relative abundance of the five most abundant bacterial phyla show changes from 3 to 20 months of life and generally remain stable after month 30 of life. **d**, The mean relative abundance of the ten most abundant bacterial pathways shows relative stability, with ABC transporters and two-component system showing the largest reduction from 3 to 20 months of life.



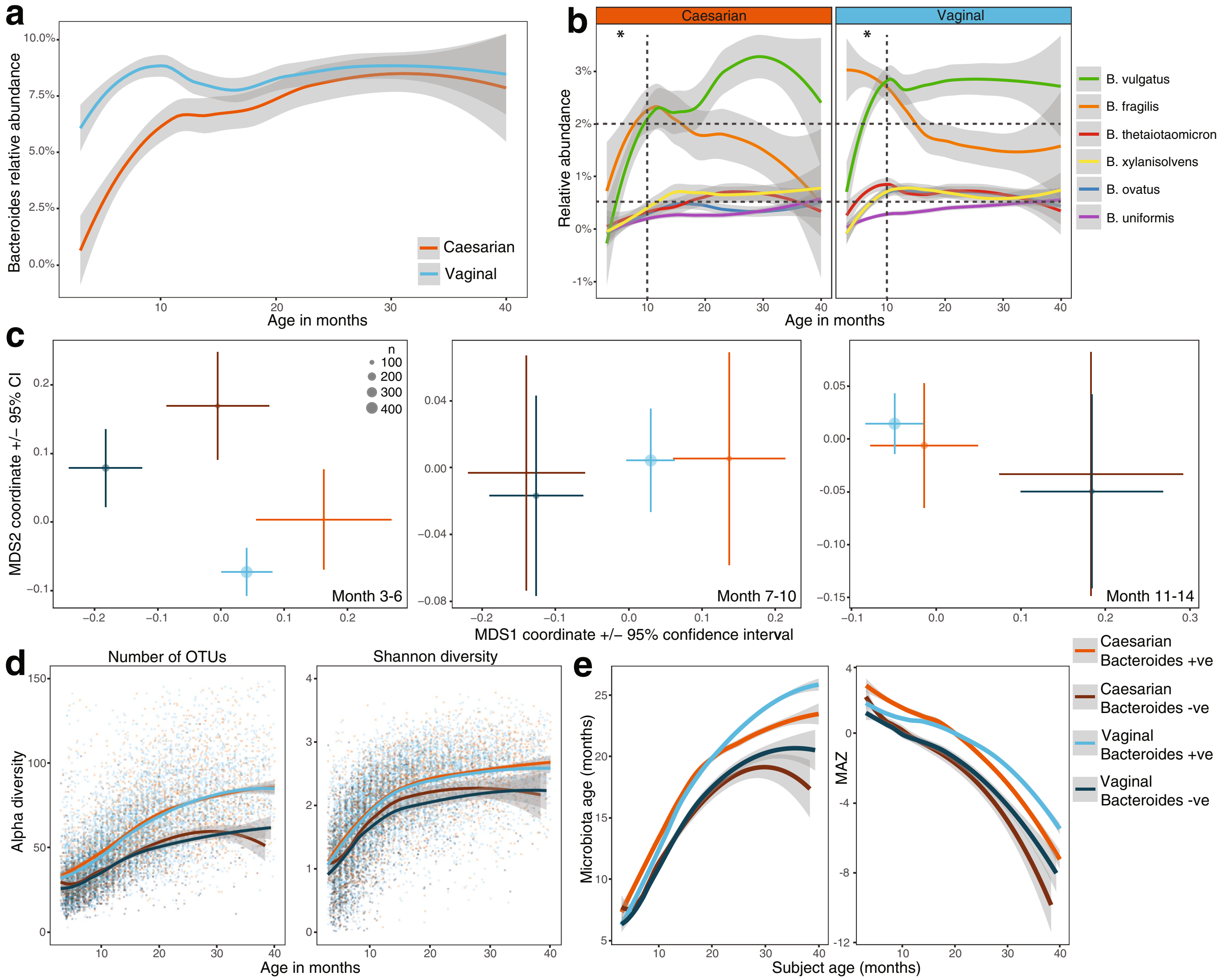
Extended Data Fig. 2 | DMM clustering of metagenomic sequencing data ($n = 10{,}867$). The entire dataset formed 18 distinct clusters based on lowest Laplace approximation. **a**, Heat map showing the relative abundance of the 25 most dominant bacterial species per each DMM cluster. **b**, Box plots showing the alpha diversity (richness and Shannon's diversity) for each DMM cluster. The centre line shows the median, the boxes cover the 25th and 75th percentiles, and the whiskers extend to the most extreme data point, which is no more than 1.5 times the length of the box away from the box. Points outside the whiskers represent outlier samples. **c**, Transition model showing the progression of samples through each DMM cluster per each time point, from months 3 to 46 of life. Dashed boxes show the three phases of microbiome progression (developmental, transitional and stable phase). Solid squares next to the labels denote the significant changes in phyla and Shannon diversity (H') per phase based on multiple linear regression. All phyla and the H' were significant in the developmental phase, two phyla and the H' were significant in the transitional phase, and no phyla or the H' were in the stable phase. Nodes and edges are sized based on the total counts. Nodes are coloured according to DMM cluster number and edges are coloured by the transition frequency. Transitions with less than 2% frequency were omitted from the plot.



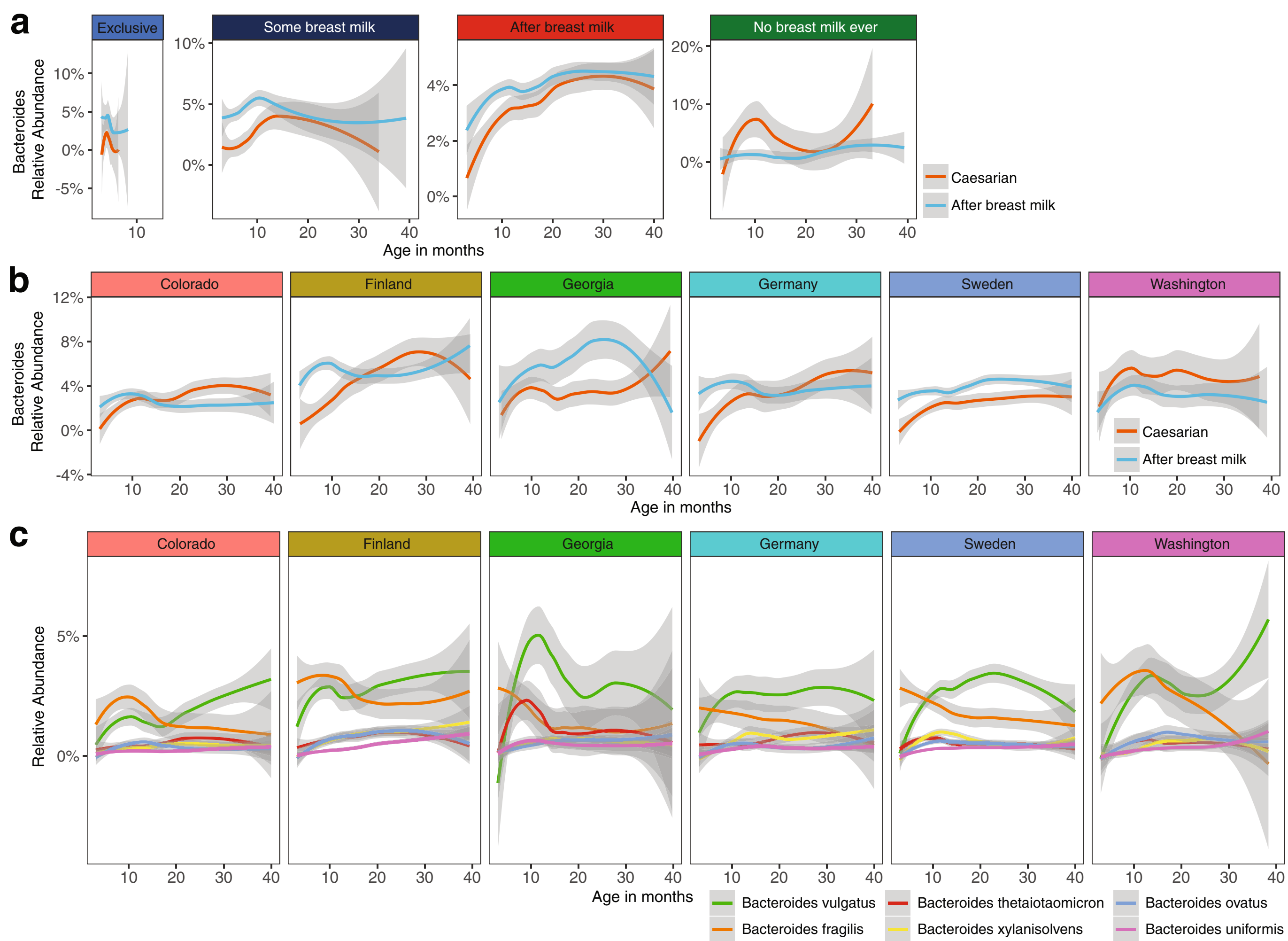
Extended Data Fig. 3 | Twenty bacterial OTUs classified by random forest regression analysis as most age discriminatory over the first 40 months of life. Rank importance of OTUs determined by applying the random forest regression to the chronological age of 150 full-term, vaginally delivered, breastfed infants ($n = 2{,}871$ stool samples). The importance of OTUs is determined by the percentage increase in mean-squared error of microbiota age prediction when the relative abundance of each OTU were randomly permuted (mean importance ± s.d., $n = 100$ replicates). These selected OTUs explained 72% of the variance (compared to 75% variance explained with all OTUs in model) and were used to define maturation of the gut microbiome by microbiota age and MAZ score. OTUs are named to the genus level and coloured based on association with life stage; blue were associated with samples collected in the first 15 months, green with samples collected between months 15 and 30, and red were with samples collected after month 30. **a**, Twenty OTUs ranked by importance to the accuracy of the model. The tenfold cross-validation error is also displayed in order of variable importance. Blue dotted line represents the 20 OTUs used in the model. **b**, Heat map of mean relative abundance of the 20 selected OTUs per month from 3 to 40 months of age.



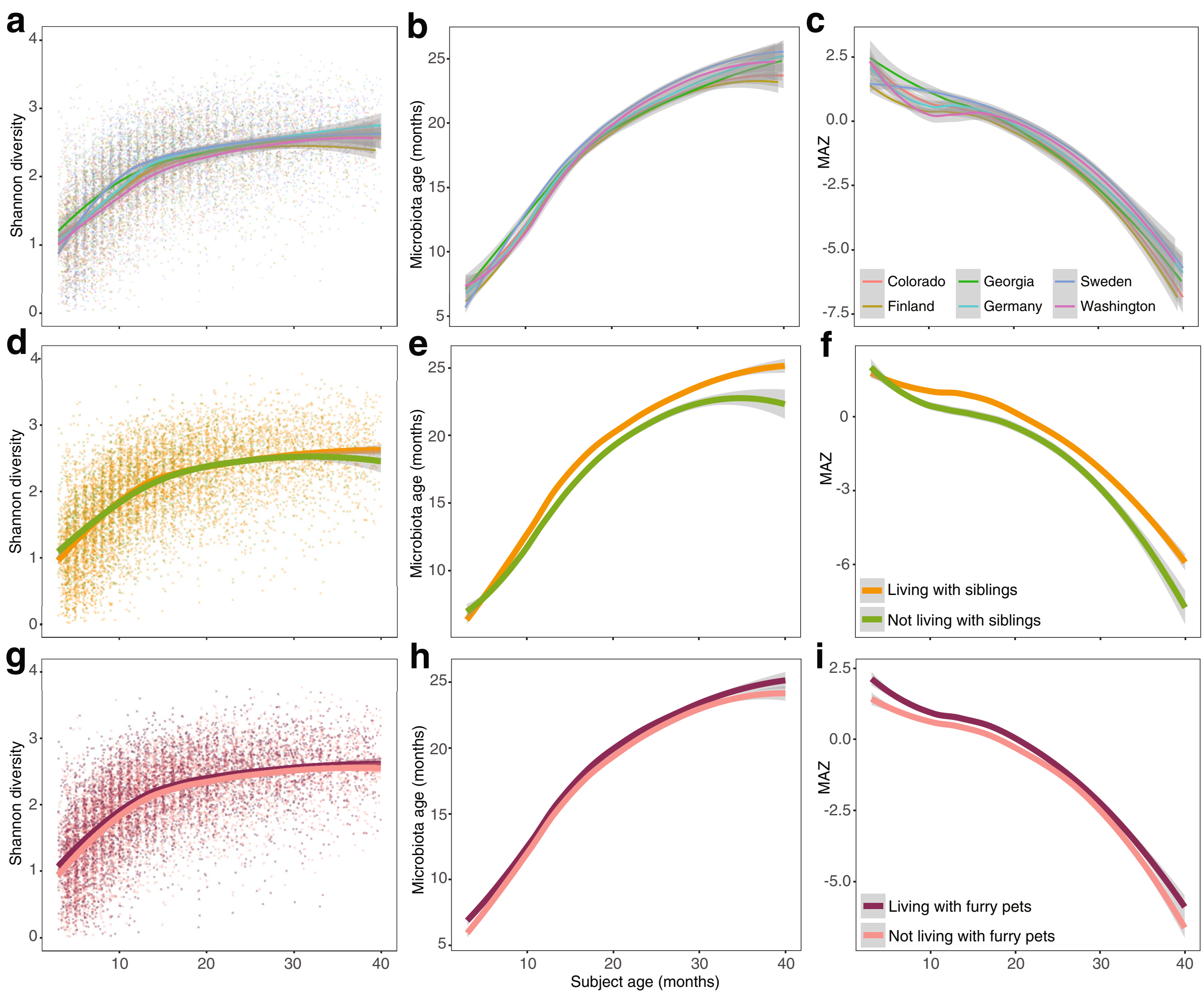
Extended Data Fig. 4 | The microbiota was not associated with the development of persistent IA and T1D. Data are based on 16S rRNA gene sequencing ($n = 11{,}717$). Analysis based on a nested 1:1 case–control cohort of equal samples. Curves show LOESS fit for the data per category, and shaded areas show permutation-based 95% confidence intervals for the fit. **a**, **b**, The number of OTUs (**a**) and the Shannon's diversity index (**b**) in the IA cohort. **c**, **d**, The number of OTUs (**c**) and Shannon's diversity (**d**) in the T1D cohort. **e**, **f**, Microbiota age (**e**) and MAZ score (**f**) in the IA cohort. **g**, **h**, Microbiota age (**g**) and MAZ score (**h**) in the T1D cohort. **i**, **j**, Forest plot showing the odds ratios for the association between the microbiome stability metrics and development of IA (**i**) and T1D (**j**). A separate conditional logistic regression was run for four time intervals: (1) birth to onset; (2) 12 months before onset; (3) 6–12 months before onset; and (4) 6 months before onset. Models were adjusted for HLA genotype, mode of delivery, duration of exclusive breastfeeding, number of antibiotic courses, and number of infectious episodes. Community states are the total number of unique clusters exhibited by an infant and state transitions are the number of transitions between clusters. No odds ratio was significantly different between cases and controls (Supplementary Table 3).



Extended Data Fig. 5 | Association of the gut microbiome with birth mode. Birth mode was significantly associated with the microbiome in months 3–6 by 16S rRNA gene sequencing and in all time points up to month 14 by metagenomic sequencing (see Supplementary Table 1). Curves show LOESS fit for the data per category, and shaded areas show permutation-based 95% confidence intervals for the fit. **a**, Longitudinal development of the *Bacteroides* genus as determined by 16S rRNA gene sequencing ($n = 11,717$). **b**, Longitudinal development of the six most abundant species within the *Bacteroides* genera as determined by metagenomic sequencing ($n = 10,867$). Grid overlay added to aid visual interpretation. **c**, NMDS ordination plots showing the mean centroid of each birth mode group stratified by *Bacteroides* positive or negative based on detection 16S rRNA gene sequencing. Plots include only the first sample obtained from a patient within a given time point for months 3–6, 7–10 and 11–14 ($n = 2,257$). Centroid size based on number of samples and the bars represent the ±95% confidence interval. **d**, Longitudinal development of the alpha diversity (richness and Shannon's diversity) with birth mode further stratified according to *Bacteroides* positive or negative ($n = 11,717$). **e**, Longitudinal development of the microbiome maturation based on the microbiota age and MAZ score against the age of the infant at sampling ($n = 11,717$). Birth mode was further stratified according to *Bacteroides* positive or negative.



Extended Data Fig. 6 | The relative abundance of *Bacteroides* stratified by breast milk and geographical location. Curves show LOESS fit for the data per category, and shaded areas show permutation-based 95% confidence intervals for the fit. **a**, **b**, *Bacteroides* genera based on 16S rRNA gene sequencing data ($n = 11{,}717$) stratified by breast milk status (**a**) and geographical location (**b**). **c**, The top 6 *Bacteroides* species based on metagenomic sequencing data ($n = 10{,}867$) stratified by geographical location.



Extended Data Fig. 7 | Environmental covariates significantly associated with the microbiome profiles. 16S rRNA gene sequencing data plotted from months 3–40 of life ($n = 11{,}717$). Curves show LOESS fit for the data per category, and shaded areas show permutation-based 95% confidence intervals for the fit. Significance determined by linear mixed-effects models in accordance with observed phases of maturation: developmental (months 3–14), transitional (months 15–30), and stable (months 31–46). Shaded lines represent the ±95% confidence interval. Longitudinal development of the Shannon's diversity index, microbiota age and MAZ score by geographical location (**a**–**c**), occurrence of household siblings (**d**–**f**), and occurrence of household furry pets (**g**–**i**).

Extended Data Table 1 | Overview of the entire analytical cohort

Month	Total cohort	3 to 6	7 to 10	11 to 14	15 to 18	19 to 22	23 to 26	27 to 30	31 to 40
Number of Subjects	903	810	800	647	543	440	336	273	220
Number of Samples	12,005	810	800	647	543	440	336	273	220
No. samples per subject (IQR)	11 (6-19)	1	1	1	1	1	1	1	1
Median age of samples in months (IQR)	13.8 (8-22.2)	4.1 (3.8-4.5)	7.7 (7.2-8.1)	11.5 (11.2-12)	15.6 (15.2-16.1)	19.7 (19.2-20.2)	23.8 (23.3-24.2)	27.7 (27.2-28.2)	31.6 (31.2-32.3)
Maternal BMI category*									
Underweight	326 (3%)	27 (3%)	25 (3%)	17 (3%)	15 (3%)	11 (3%)	9 (3%)	6 (2%)	5 (2%)
Normal	7209 (60%)	478 (59%)	479 (60%)	379 (59%)	328 (60%)	265 (60%)	204 (61%)	165 (60%)	140 (64%)
Overweight	2991 (25%)	197 (24%)	189 (24%)	162 (25%)	132 (24%)	114 (26%)	92 (27%)	75 (27%)	55 (25%)
Obese	1386 (12%)	100 (12%)	101 (13%)	83 (13%)	64 (12%)	47 (11%)	28 (8%)	25 (9%)	18 (8%)
NA	93 (1%)	8 (1%)	6 (1%)	6 (1%)	4 (1%)	3 (1%)	3 (1%)	2 (1%)	(0%)
Maternal preg. weight gain*									
Non-excess	8497 (71%)	584 (72%)	579 (72%)	473 (73%)	386 (71%)	308 (70%)	236 (70%)	193 (71%)	151 (69%)
Excess	3415 (28%)	218 (27%)	215 (27%)	168 (26%)	153 (28%)	129 (29%)	97 (29%)	78 (29%)	67 (30%)
NA	93 (1%)	8 (1%)	6 (1%)	6 (1%)	4 (1%)	3 (1%)	3 (1%)	2 (1%)	2 (1%)
Maternal diabetes medication*									
Insulin	977 (8%)	66 (8%)	64 (8%)	56 (9%)	41 (8%)	28 (6%)	23 (7%)	23 (8%)	19 (9%)
Metformin and insulin	38 (0%)	3 (0%)	2 (0%)	1 (0%)	1 (0%)	1 (0%)	1 (0%)	1 (0%)	1 (0%)
Glyburide	126 (1%)	5 (1%)	6 (1%)	6 (1%)	5 (1%)	3 (1%)	4 (1%)	4 (1%)	3 (1%)
None	10864 (90%)	736 (91%)	728 (90%)	584 (90%)	496 (91%)	400 (91%)	304 (90%)	245 (90%)	197 (90%)
Maternal diabetes*									
Gestational	779 (6%)	56 (7%)	53 (7%)	43 (7%)	36 (7%)	29 (7%)	21 (6%)	16 (6%)	16 (7%)
None	10060 (84%)	673 (83%)	672 (84%)	535 (83%)	458 (84%)	368 (84%)	280 (83%)	229 (84%)	185 (84%)
T1D	839 (7%)	53 (7%)	54 (7%)	49 (8%)	35 (6%)	29 (7%)	25 (7%)	22 (8%)	16 (7%)
T2D	30 (0%)	3 (0%)	3 (0%)	2 (0%)	2 (0%)	1 (0%)	0 (0%)	0 (0%)	0 (0%)
NA	297 (2%)	25 (3%)	18 (2%)	18 (3%)	12 (2%)	13 (3%)	10 (3%)	6 (2%)	3 (1%)
Maternal preeclampsia*									
Yes	445 (4%)	35 (4%)	32 (4%)	23 (4%)	21 (4%)	17 (4%)	11 (3%)	9 (3%)	9 (4%)
No	11439 (95%)	765 (94%)	757 (95%)	618 (96%)	516 (95%)	418 (95%)	322 (96%)	262 (96%)	210 (95%)
NA	121 (1%)	10 (1%)	11 (1%)	6 (1%)	6 (1%)	5 (1%)	3 (1%)	2 (1%)	1 (0%)
Maternal probiotic*									
Yes	584 (5%)	43 (5%)	42 (5%)	30 (5%)	24 (4%)	21 (5%)	17 (5%)	14 (5%)	9 (4%)
No	11421 (95%)	767 (95%)	758 (95%)	617 (95%)	519 (96%)	419 (95%)	319 (95%)	259 (95%)	211 (96%)
Birth mode									
Caesarian	3319 (28%)	196 (24%)	200 (25%)	170 (26%)	146 (27%)	121 (28%)	88 (26%)	84 (31%)	67 (30%)
Vaginal	8686 (72%)	614 (76%)	600 (75%)	477 (74%)	397 (73%)	319 (73%)	248 (74%)	189 (69%)	153 (70%)
Preterm (<37 weeks)									
Yes	723 (6%)	39 (5%)	43 (5%)	35 (5%)	30 (6%)	29 (7%)	20 (6%)	16 (6%)	16 (7%)
No	11282 (94%)	771 (95%)	757 (95%)	612 (95%)	513 (94%)	411 (93%)	316 (94%)	257 (94%)	204 (93%)
Geographical location									
Colorado	2079 (17%)	122 (15%)	120 (15%)	99 (15%)	84 (15%)	57 (13%)	45 (13%)	39 (14%)	30 (14%)
Finland	2833 (24%)	223 (28%)	215 (27%)	163 (25%)	131 (24%)	107 (24%)	85 (25%)	57 (21%)	40 (18%)
Georgia	872 (7%)	46 (6%)	53 (7%)	50 (8%)	48 (9%)	33 (8%)	23 (7%)	22 (8%)	18 (8%)
Germany	1219 (10%)	86 (11%)	78 (10%)	60 (9%)	46 (8%)	46 (10%)	34 (10%)	28 (10%)	16 (7%)
Sweden	4024 (34%)	260 (32%)	262 (33%)	216 (33%)	190 (35%)	150 (34%)	119 (35%)	102 (37%)	99 (45%)
Washington	978 (8%)	73 (9%)	72 (9%)	59 (9%)	44 (8%)	47 (11%)	30 (9%)	25 (9%)	17 (8%)
Sex									
Male	6455 (54%)	497 (61%)	438 (55%)	357 (55%)	297 (55%)	230 (52%)	175 (52%)	146 (53%)	119 (54%)
Female	5550 (46%)	372 (46%)	362 (45%)	290 (45%)	246 (45%)	210 (48%)	161 (48%)	127 (47%)	101 (46%)
Race/Ethnicity**									
African Americans	24 (0%)	2 (0%)	2 (0%)	1 (0%)	1 (0%)	1 (0%)	1 (0%)	2 (1%)	0 (0%)
Hispanic	445 (4%)	32 (4%)	32 (4%)	29 (4%)	22 (4%)	12 (3%)	10 (3%)	4 (1%)	3 (1%)
White Non-Hispanic	7391 (62%)	512 (63%)	497 (62%)	395 (61%)	322 (59%)	272 (62%)	201 (60%)	161 (59%)	199 (90%)
All Other Races	128 (1%)	6 (1%)	8 (1%)	6 (1%)	8 (1%)	4 (1%)	5 (1%)	5 (2%)	2 (1%)
NA	4017 (33%)	258 (32%)	261 (33%)	216 (33%)	190 (35%)	151 (34%)	119 (35%)	101 (37%)	16 (37%)
Breast milk status									
No breast milk ever	220 (2%)	13 (2%)	15 (2%)	12 (2%)	11 (2%)	7 (2%)	7 (2%)	8 (3%)	3 (1%)
Exclusive	240 (2%)	136 (17%)	4 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Some breast milk	3048 (25%)	445 (55%)	417 (52%)	170 (26%)	52 (10%)	27 (6%)	10 (3%)	7 (3%)	5 (2%)
After breast milk	8497 (71%)	216 (27%)	364 (46%)	465 (72%)	480 (88%)	406 (92%)	319 (95%)	258 (95%)	212 (96%)
Solid food status									
Before	602 (5%)	373 (46%)	4 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
After	11403 (95%)	437 (54%)	796 (100%)	647 (100%)	543 (100%)	440 (100%)	336 (100%)	273 (100%)	220 (100%)
Probiotic									
Never	8017 (67%)	516 (64%)	508 (64%)	432 (67%)	363 (67%)	288 (65%)	222 (66%)	184 (67%)	149 (68%)
Before	1183 (10%)	155 (19%)	117 (15%)	63 (10%)	45 (8%)	27 (6%)	18 (5%)	9 (3%)	10 (5%)
During	2296 (19%)	122 (15%)	153 (19%)	136 (21%)	115 (21%)	107 (24%)	78 (23%)	57 (21%)	42 (19%)
After	509 (4%)	17 (2%)	22 (3%)	16 (2%)	20 (4%)	18 (4%)	18 (5%)	23 (8%)	19 (9%)
Vitamin D									
Never	692 (6%)	48 (6%)	46 (6%)	39 (6%)	30 (6%)	21 (5%)	17 (5%)	13 (5%)	9 (4%)
During	11197 (93%)	756 (93%)	747 (93%)	602 (93%)	508 (94%)	413 (94%)	313 (93%)	257 (94%)	208 (95%)
After	116 (1%)	6 (1%)	7 (1%)	6 (1%)	5 (1%)	6 (1%)	6 (2%)	3 (1%)	3 (1%)
Antibiotics									
Never	3796 (32%)	241 (30%)	240 (30%)	198 (31%)	166 (31%)	138 (31%)	106 (32%)	83 (30%)	78 (35%)
Before	2358 (20%)	415 (51%)	287 (36%)	152 (23%)	63 (12%)	11 (3%)	0 (0%)	0 (0%)	0 (0%)
During	288 (2%)	19 (2%)	26 (3%)	24 (4%)	18 (3%)	12 (3%)	0 (0%)	0 (0%)	0 (0%)
After	5563 (46%)	135 (17%)	247 (31%)	273 (42%)	296 (55%)	279 (63%)	230 (68%)	190 (70%)	142 (65%)
Household siblings									
Never	2187 (18%)	168 (21%)	164 (21%)	131 (20%)	108 (20%)	79 (18%)	53 (16%)	33 (12%)	25 (11%)
Before	2890 (24%)	167 (21%)	171 (21%)	147 (23%)	131 (24%)	115 (26%)	87 (26%)	76 (28%)	58 (26%)
During	6877 (57%)	468 (58%)	460 (58%)	367 (57%)	303 (56%)	245 (56%)	194 (58%)	163 (60%)	136 (62%)
After	7 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (1%)	1 (0%)	1 (0%)
NA	44 (0%)	7 (1%)	5 (1%)	2 (0%)	1 (0%)	1 (0%)	0 (0%)	0 (0%)	0 (0%)
Household flurry pets									
Never	6001 (50%)	418 (52%)	414 (52%)	329 (51%)	275 (51%)	227 (52%)	166 (49%)	133 (49%)	107 (49%)
Before	779 (6%)	50 (6%)	45 (6%)	39 (6%)	36 (7%)	30 (7%)	25 (7%)	20 (7%)	15 (7%)
During	5160 (43%)	336 (41%)	337 (42%)	278 (43%)	232 (43%)	183 (42%)	141 (42%)	116 (42%)	96 (44%)
After	40 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (1%)	4 (1%)	2 (1%)
NA	25 (0%)	6 (1%)	4 (1%)	1 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Lives on farm with animals									
Never	10385 (87%)	692 (85%)	688 (86%)	559 (86%)	474 (87%)	379 (86%)	289 (86%)	238 (87%)	193 (88%)
Before	478 (4%)	35 (4%)	31 (4%)	26 (4%)	25 (5%)	19 (4%)	14 (4%)	10 (4%)	6 (3%)
During	1085 (9%)	77 (10%)	77 (10%)	60 (9%)	43 (8%)	42 (10%)	31 (9%)	22 (8%)	19 (9%)
After	32 (0%)	0 (0%)	0 (0%)	1 (0%)	1 (0%)	0 (0%)	2 (1%)	3 (1%)	2 (1%)
NA	25 (0%)	6 (1%)	4 (1%)	1 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Daycare exposure									
Yes	4600 (38%)	49 (6%)	115 (14%)	139 (21%)	237 (44%)	263 (60%)	227 (68%)	194 (71%)	165 (75%)
No	7405 (62%)	761 (94%)	685 (86%)	508 (79%)	306 (56%)	177 (40%)	109 (32%)	79 (29%)	55 (25%)
Celiac disease									
Yes	506 (4%)	27 (3%)	28 (4%)	25 (4%)	22 (4%)	19 (4%)	12 (4%)	13 (5%)	10 (5%)
No	11499 (96%)	783 (97%)	772 (97%)	622 (96%)	521 (96%)	421 (96%)	324 (96%)	260 (95%)	210 (95%)
Any chronic disease/disorder									
Never	8937 (74%)	622 (77%)	604 (76%)	486 (75%)	411 (76%)	333 (76%)	247 (74%)	195 (71%)	161 (73%)
Before	1815 (15%)	154 (19%)	136 (17%)	102 (16%)	72 (13%)	54 (12%)	37 (11%)	32 (12%)	20 (9%)
During	1212 (10%)	33 (4%)	58 (7%)	57 (9%)	59 (11%)	52 (12%)	51 (15%)	43 (16%)	37 (17%)
After	41 (0%)	1 (0%)	2 (0%)	2 (0%)	1 (0%)	1 (0%)	1 (0%)	3 (1%)	2 (1%)
Any acute disease/disorder									
Never	17 (0%)	4 (0%)	3 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Before	1469 (12%)	599 (74%)	121 (15%)	18 (3%)	3 (1%)	1 (0%)	0 (0%)	0 (0%)	0 (0%)
During	5107 (43%)	165 (20%)	321 (40%)	308 (48%)	265 (49%)	191 (43%)	151 (45%)	135 (49%)	96 (44%)
After	5412 (45%)	42 (5%)	355 (44%)	321 (50%)	275 (51%)	248 (56%)	185 (55%)	138 (51%)	124 (56%)

IQR, interquartile range; NA, not available.
*Maternal variables relate to measurements obtained during pregnancy.
**Race/ethnicity was only available for the US sites.

OPEN
https://doi.org/10.1038/s41586-018-0620-2

The human gut microbiome in early-onset type 1 diabetes from the TEDDY study

Tommi Vatanen[1]*, Eric A. Franzosa[1,2], Randall Schwager[2], Surya Tripathi[1], Timothy D. Arthur[1], Kendra Vehik[3], Åke Lernmark[4], William A. Hagopian[5], Marian J. Rewers[6], Jin-Xiong She[7], Jorma Toppari[8,9], Anette-G. Ziegler[10,11,12], Beena Akolkar[13], Jeffrey P. Krischer[3], Christopher J. Stewart[14,15], Nadim J. Ajami[14], Joseph F. Petrosino[14], Dirk Gevers[1,19], Harri Lähdesmäki[16], Hera Vlamakis[1], Curtis Huttenhower[1,2,20]* & Ramnik J. Xavier[1,17,18,20]*

Type 1 diabetes (T1D) is an autoimmune disease that targets pancreatic islet beta cells and incorporates genetic and environmental factors[1], including complex genetic elements[2], patient exposures[3] and the gut microbiome[4]. Viral infections[5] and broader gut dysbioses[6] have been identified as potential causes or contributing factors; however, human studies have not yet identified microbial compositional or functional triggers that are predictive of islet autoimmunity or T1D. Here we analyse 10,913 metagenomes in stool samples from 783 mostly white, non-Hispanic children. The samples were collected monthly from three months of age until the clinical end point (islet autoimmunity or T1D) in the The Environmental Determinants of Diabetes in the Young (TEDDY) study, to characterize the natural history of the early gut microbiome in connection to islet autoimmunity, T1D diagnosis, and other common early life events such as antibiotic treatments and probiotics. The microbiomes of control children contained more genes that were related to fermentation and the biosynthesis of short-chain fatty acids, but these were not consistently associated with particular taxa across geographically diverse clinical centres, suggesting that microbial factors associated with T1D are taxonomically diffuse but functionally more coherent. When we investigated the broader establishment and development of the infant microbiome, both taxonomic and functional profiles were dynamic and highly individualized, and dominated in the first year of life by one of three largely exclusive *Bifidobacterium* species (*B. bifidum*, *B. breve* or *B. longum*) or by the phylum Proteobacteria. In particular, the strain-specific carriage of genes for the utilization of human milk oligosaccharide within a subset of *B. longum* was present specifically in breast-fed infants. These analyses of TEDDY gut metagenomes provide, to our knowledge, the largest and most detailed longitudinal functional profile of the developing gut microbiome in relation to islet autoimmunity, T1D and other early childhood events. Together with existing evidence from human cohorts[7,8] and a T1D mouse model[9], these data support the protective effects of short-chain fatty acids in early-onset human T1D.

Recent literature has linked several facets of gut health with the onset of T1D in humans and rodent models[4,6,10]. Altered intestinal microbiota in connection to T1D has been reported in Finnish[7,8,11,12], German[13], Italian[14], Mexican[15], American (Colorado)[16] and Turkish[17] children. Common findings include increased numbers of *Bacteroides* species, and deficiency of bacteria that produce short-chain fatty acids (SCFAs)[7,8] in cases of T1D or islet autoimmunity (IA)[8,11,15,18]. Corroborating these findings, decreased levels of SCFA-producing bacteria were found in adults with type 2 diabetes (T2D)[19]. In addition, increased intestinal permeability[14] and decreased microbial diversity[12] after IA but before T1D diagnosis have been reported. Studies using the nonobese diabetic (NOD) mouse model have determined immune mechanisms that mediate the protective effects of SCFAs[9] and the microbiome-linked sex bias in autoimmunity[20]. NOD mice fed specialized diets resulting in high bacterial release of the SCFAs acetate and butyrate were almost completely protected from T1D[9]. A study in a streptozotocin-induced T1D mouse model demonstrated that bacterial products recognized in pancreatic lymph nodes contribute to pathogenesis[21].

Even in the absence of immune perturbation, the first few weeks, months and years of life represent a unique human microbial environment that has only recently been detailed[22,23]. Infants have a markedly different gut microbial profile from adults, characterized by a distinct taxonomic profile, greater proportion of aerobic energy harvest metabolism, and more extreme dynamic change[24]. These differences gradually fade over the first few years of life, particularly in response to the introduction of solid food, and individual microbial developmental trajectories are influenced by environment, delivery mode, breast (versus formula) feeding, and antibiotics[25–27]. Most studies that address the development of the gut microbiome, both generally and in association with T1D, have used gene analysis of 16S rRNA, which leaves open the question of functional and strain-specific differences that are not easily detected by this technology that might contribute to disease pathogenesis[12].

Bridging this gap is one goal of the The Environmental Determinants of Diabetes in the Young (TEDDY) study, a prospective study that aims to identify environmental causes of T1D[28]. It includes six clinical research centres in the United States (Colorado, Georgia/Florida and Washington) and Europe (Finland, Germany and Sweden), which together have recruited several thousand newborns with a genetic predisposition for T1D or first-degree relative(s) with T1D. This has enabled the TEDDY study to collect a range of biospecimens, including monthly stool samples starting at three months of age, coupled with extensive clinical and personal data such as diet, illnesses, medications and other life experiences. To characterize microbial, environmental, genetic, immunological and additional contributors to the development

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [3]Health Informatics Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA. [4]Department of Clinical Sciences, Lund University/CRC, Skåne University Hospital SUS, Malmo, Sweden. [5]Pacific Northwest Research Institute, Seattle, WA, USA. [6]Barbara Davis Center for Childhood Diabetes, University of Colorado, Aurora, CO, USA. [7]Center for Biotechnology and Genomic Medicine, Medical College of Georgia, Augusta University, Augusta, GA, USA. [8]Department of Pediatrics, Turku University Hospital, Turku, Finland. [9]Institute of Biomedicine, Research Centre for Integrative Physiology and Pharmacology, University of Turku, Turku, Finland. [10]Institute of Diabetes Research, Helmholtz Zentrum München, Munich, Germany. [11]Forschergruppe Diabetes, Technische Universität München, Klinikum Rechts der Isar, Munich, Germany. [12]Forschergruppe Diabetes e.V. at Helmholtz Zentrum München, Munich, Germany. [13]National Institute of Diabetes & Digestive & Kidney Diseases, Bethesda, MD, USA. [14]Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA. [15]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK. [16]Department of Computer Science, Aalto University, Espoo, Finland. [17]Gastrointestinal Unit, Center for the Study of Inflammatory Bowel Disease, and Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [18]Center for Microbiome Informatics and Therapeutics, MIT, Cambridge, MA, USA. [19]Present address: Janssen Human Microbiome Institute, Janssen Research and Development, Cambridge, MA, USA. [20]These authors jointly supervised this work: Curtis Huttenhower, Ramnik J. Xavier. *e-mail: vatanen@broadinstitute.org; chuttenh@hsph.harvard.edu; xavier@molbio.mgh.harvard.edu

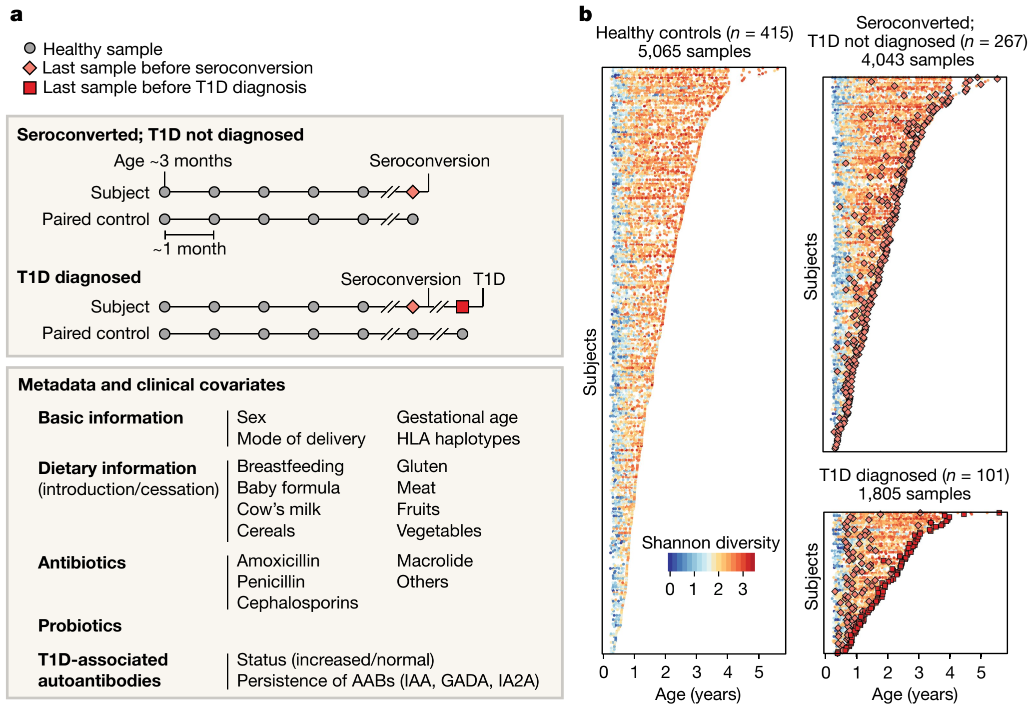


Fig. 1 | More than 10,000 longitudinal gut metagenomes from the TEDDY T1D cohort. We analysed 10,913 metagenomes collected longitudinally from 783 children (415 controls, 267 seroconverters, and 101 diagnosed with T1D) approximately monthly over the first five years of life. **a**, Subjects were recruited at six clinical centres (Finland, Sweden, Germany, Washington, Georgia and Colorado). Primary end points were seroconversion (defined as persistent confirmed IA) and T1D diagnosis. Additional metadata analysed for subjects and samples included the status of breastfeeding, birth mode, probiotics, antibiotics, formula feeding, and other dietary covariates. **b**, Overview of stool samples collected and microbiome development as summarized by Shannon's alpha diversity and stratified by end point. Median number of samples per individual $n = 12$ (healthy controls $n = 10$, seroconverters $n = 13$, T1D cases $n = 16$).

of T1D, the TEDDY study group further assembled nested case–control studies for IA ($n = 418$ case–control pairs) and T1D ($n = 114$)[29]. Case–control pairs were matched by clinical centre, sex and family history of T1D, which are all known confounding factors for T1D susceptibility and microbiome composition.

Here, we assessed 783 children followed from three months to up to five years of age from six clinical centres in four countries (Finland, Germany, Sweden and the United States) who either progressed to persistent IA or T1D or were matched as controls (Fig. 1a, b, Extended Data Table 1). Stool samples were collected, on average, monthly starting at three months of age and continuing until the clinical end point (IA or T1D). This study focused solely on analysing metagenomic sequencing data ($n = 10{,}903$ samples, $n = 783$ subjects), while a companion paper by Stewart et al.[30] interrogated corresponding 16S rRNA amplicon sequencing information.

We first investigated the taxonomic composition of early gut metagenomes at the species level. Principal coordinate analysis ordination of Bray–Curtis beta diversities showed a strong longitudinal gradient and marked heterogeneity among the earliest samples (Fig. 2a, Extended Data Fig. 1a–k, Supplementary Note 1). Permutational analysis of variance (ANOVA) of Bray–Curtis beta diversities indicated that inter-subject differences explained 35% of microbial taxonomic variation (permutation test, $P < 0.001$, 1,000 permutations), followed by age at stool sampling at roughly 4% of variance ($P < 0.001$). Using cross-sectional analysis to test for associations between taxonomic beta diversities and other collected metadata, we found that in addition to subject ID and age, geographical location and breastfeeding had strong and systematic effects on the composition of the microbial community (Supplementary Table 1, Extended Data Fig. 2a–d, Supplementary Note 1). To investigate the stability and individuality of the microbial profiles further, we compared intra- and inter-subject Bray–Curtis beta diversities. The gap between individual stability and similarity within or across clinical centres was largest at the beginning of the sampling period, indicating that the children had particularly dissimilar microbiota during these early months (Fig. 2b, Supplementary Note 1). Finally, we tested microbial alpha diversity (Shannon's diversity index) of taxonomic profiles for associations with collected metadata, and found that the cessation of breastfeeding had the largest effect (ANOVA, partial $\eta^2 = 0.053$) in the accrual of alpha diversity in early life (Supplementary Table 2, Extended Data Fig. 3a–e, Supplementary Note 1).

We next investigated the effects of antibiotics on the early life microbiome. Courses of oral antibiotics disrupted microbial stability, with a larger effect in the earliest comparisons (Fig. 2c, Extended Data Fig. 4a–f, Extended Data Table 2, Supplementary Note 2). Previous studies have found *Bifidobacterium* species to be especially vulnerable to antibiotics[31,32], leading us to investigate how antibiotic perturbations influenced these common dominant members of the early gut. Comparing microbial relative abundances before and after antibiotics (assuming that the given species was present in the preceding sample), we saw a decrease in the abundances of the *Bifidobacterium* members *B. bifidum*, *B. pseudocatenulatum*, *B. adolescentis*, *B. dentium* and *B. catenulatum*, whereas *B. longum* and *B. breve* did not systematically decline owing to antibiotics (Fig. 2d), suggesting that certain *Bifidobacterium* species are particularly susceptible to out-competition by other community members after depletion by antibiotics. Given their dominance in the typical developing gut microbiota and finely tuned balance of metabolic interactions with breast milk, this finding underscores the importance of approaching antibiotic prescriptions in early childhood with care, especially during breastfeeding.

Accompanying our taxonomic profiling, functional profiling of these metagenomes suggested the development of a consistent microbial functional core during infancy, with a smaller subject-specific variable functional pool (Extended Data Fig. 5a, b, Supplementary Note 3). As in most microbial community studies[33], microbial gene families of uncharacterized function made up a substantial fraction of these

profiles, averaging roughly 50% based on Gene Ontology[34] annotations (Extended Data Fig. 5c) and more than 90% based on more functionally specific MetaCyc pathways (Extended Data Fig. 5d). We observed an increasing longitudinal trend in the proportion of unmapped reads (Extended Data Fig. 5e, Pearson's $r=0.318$, $P<2.2\times10^{-16}$). However, within the reads that mapped to either microbial pangenomes or known protein sequences (the proportion of which decreased with age), we saw an increase in the proportion of reads with MetaCyc annotation, mainly during the first year (Extended Data Fig. 5f, Pearson $r=0.391$, $P<2.2\times10^{-16}$). This suggests that although the early life microbiome is relatively well-covered by current microbial reference genomes, less functional and biochemical characterization has been carried out on gene families within these microorganisms, which will thus particularly benefit from future work.

In addition to broadly conserved and subject-specific functions, we identified a range of microbial metabolic enzymes that consistently increased or decreased in abundance over the first year of life, paralleling shifts in community structure and infant diet (Fig. 3, Supplementary Note 3, Supplementary Table 3). For example, the enzyme L-lactate dehydrogenase (1.1.1.27), which is well-characterized in Bifidobacteria for its role in milk fermentation[35], was among the most consistently declining enzymes over this period, notably coinciding with the cessation of breastfeeding in many infants (from 73% breastfed at month 3 to 28% at year 1). Conversely, the enzyme transketolase (2.2.1.1), which has been implicated previously[36] in the metabolism of fibre, was among the most consistently increasing enzymes, which also coincided with increased incorporation of solid food (a component of 53% of infants' diets at month 3 versus 100% at year 1). Hence, these notable changes in community functional potential highlight the unique metabolic environment of the early infant gut, and the subsequent transition to a more adult-like gut microbiome that is adapted to variable, fermentative energy sources.

Combining taxonomic and functional profiles to test for differences between cases and controls, we used linear mixed-effects modelling and identified a relatively small number of individual taxonomic and functional features that were associated with case–control outcome (Supplementary Table 4), most with borderline statistical significance (false discovery rate (FDR) corrected q-values indicated below). We confirmed separation between cases and controls by random forest classifiers (Extended Data Fig. 6a, b, Supplementary Note 4). In the IA case–control cohort, healthy controls contained higher levels of *Lactobacillus rhamnosus* ($q=0.055$), supporting protection against IA by early probiotic supplementation[37] (Extended Data Fig. 6c, d, Supplementary Note 5). IA controls also had more *Bifidobacterium dentium* ($q=0.054$), whereas IA cases had on average higher abundance of *Streptococcus* group *mitis/oralis/pneumoniae* species ($q=0.11$). In T1D case–control comparisons, controls had higher levels of *Streptococcus thermophilus* ($q=0.078$) and *Lactococcus lactis* ($q=0.094$) species, both common in dairy products, whereas cases contained higher levels of species such as *Bifidobacterium pseudocatenulatum* ($q=0.078$), *Roseburia hominis* ($q=0.11$) and *Alistipes shahii* ($q=0.14$). Even though our modelling approach controlled for regional differences in clinical centres, we found additional but often weak associations with outcome in some clinical centres when tested separately (Supplementary Table 4). Finnish IA cases had more *Streptococcus* group *mitis/oralis/pneumoniae* species ($q=0.0008$), IA controls from Colorado had more *Streptococcus thermophilus* ($q=0.0059$), and Swedish IA cases contained more *Bacteroides vulgatus* ($q=0.090$).

Pathways with the highest statistical significance in case–control comparisons were related to bacterial fermentation (Supplementary Table 4). The superpathway of fermentation (MetaCyc identifier PWY4LZ-257) was increased in controls in the T1D cohort ($q=0.019$) and Finnish IA cohort ($q=0.049$). SCFAs such as butyrate, acetate and propionate are common by-products of bacterial fermentation, and butyrate and acetate protected NOD mice against T1D[9]. Consistently, we observed that several bacterial pathways that contribute to the biosynthesis of short-chain fatty acids were increased in healthy controls.

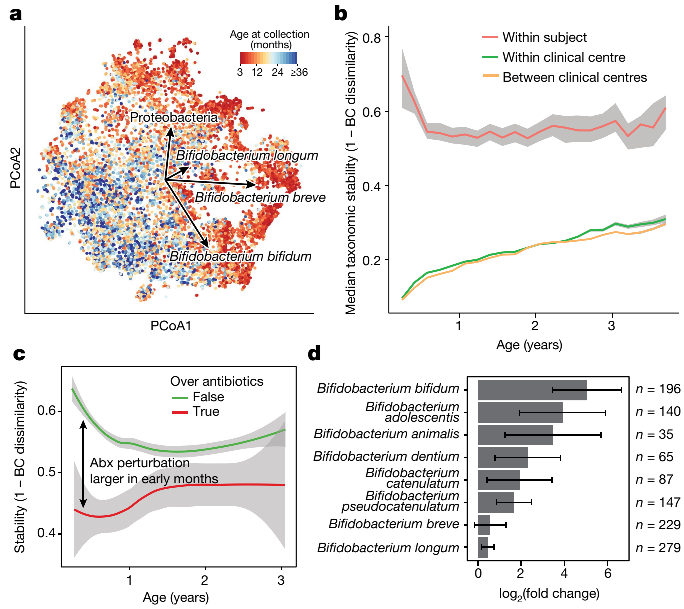


Fig. 2 | The early gut microbiome is characterized by early heterogeneity of *Bifidobacterium* species and individualized accrual of taxa over time. a, Principal coordinate analysis (PCoA) ordination of microbial beta diversities ($n=10{,}913$ samples), measured by Bray–Curtis dissimilarity. Arrows show the weighted averages of key taxonomic groups. **b**, Microbiota stability, measured by Bray–Curtis (BC) dissimilarity ($n=10{,}750$ samples) in three-month time windows, over two-month increments, stratified into three groups: within subject, within clinical centre, and between clinical centres. Lines show median values per time window. Shaded area denotes the estimated 99% confidence interval. Gut microbial communities were highly individual. **c**, Influence of antibiotic (Abx) courses on microbial stability, measured by Bray–Curtis dissimilarity over consecutive stool samples (<50 days apart) from the same individual during the first three years of life, and stratified by whether antibiotics were given between the two samples ($n=654$ observations with antibiotics, $n=6{,}734$ observations without antibiotics). Curves show locally weighted scatterplot smoothing (LOESS) for the data per category. Shaded areas show permutation-based 95% confidence intervals for the fit. **d**, Decreases in the most common *Bifidobacterium* species in connection to oral antibiotic treatments. Fold change was measured between consecutive samples with an antibiotic course between them, given that the species in question was present in the first of the two samples. Sample size per species (n) indicates the number of sample pairs in which the species in question was present in the sample before the antibiotic treatment. Bars show bootstrapped mean $\log_2$(fold change) (that is, decrease), and error bars denote s.d. ($n=1{,}000$ bootstrap samples).

Among pathways involved in butyrate production, the degradation of L-arginine, putrescine and 4-aminobutanoate (ARGDEG-PWY) superpathway was increased in T1D controls cohort-wide ($q=0.043$), whereas the fermentation of acetyl coenzyme A to butanoate (PWY-5676) was more abundant in the Finnish T1D controls ($q=0.053$). The degradation of acetylene (P161-PWY), which contributes to acetate production, was increased in T1D controls cohort-wide ($q=0.14$), and the degradation of L-1,2-propanediol (PWY-7013), which is involved in propionate biosynthesis, was higher in the German T1D controls ($q=0.019$). These findings support existing evidence for the protective effects of SCFAs in human T1D[7,8] and T2D[19] cohorts and the NOD mouse model[9].

As reflected by the community-level analyses, human milk with its pro- and prebiotic functions is one of the main factors that determine the community composition of the infant gut microbiome. *Bifidobacterium longum* subsp. *infantis* is a particularly versatile degrader of human milk oligosaccharide (HMO) that is often found in stool samples collected during breastfeeding[38]. By following the families representing genes in the *B. longum* subsp. *infantis* HMO gene cluster[39,40] in our data, we found that an additional 30 bacterial species

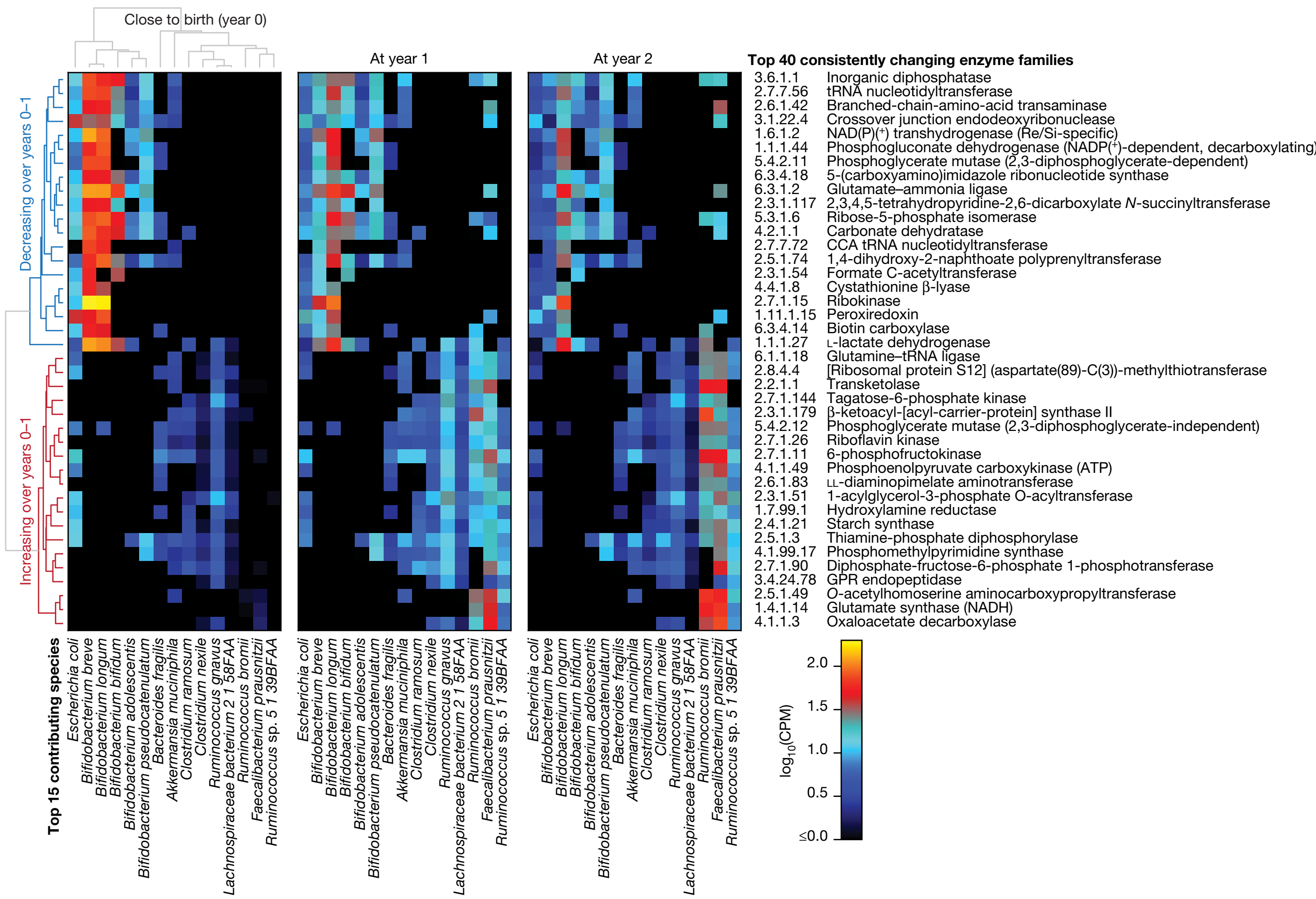


Fig. 3 | Consistent changes in enzymatic content of the gut microbiome in early life. We identified enzyme families (level-4 Enzyme Commission (EC) categories) that exhibited the most consistent within-subject changes in total community abundance between the ages of 3 months and 1 year. The top 20 most consistent increases or decreases are presented and stratified according to their top 15 contributing species. Heat map values reflect the mean contribution of each species to each enzyme over samples ($n = 733$ at 3 months; 675 at 1 year; and 382 at 2 years). Values reflect units of copies per million (CPM) normalized to total read depth (including unmapped reads and reads mapped to gene families lacking EC annotation). Rows (enzymes) and columns (species) are clustered according to Spearman correlation at 3 months; subsequent years are ordered according to clustering at 3 months.

carried at least one homologue with more than 50% sequence identity to one or more HMO utilization genes (Supplementary Table 5). As expected, many Bifidobacteria carried several homologues, but surprisingly three *Enterococcus* species (*E. casseliflavus*, *E. faecalis* and *E. faecium*) also carried seven or more homologues (Supplementary Table 5).

To identify strain-level adaptation similar to *B. longum* subsp. *infantis*, we further examined whether any of these genes showed contrasting prevalence between samples collected during breastfeeding and after weaning, given that the carrier species itself was present. In total, 41 gene families were observed more often during breastfeeding (Supplementary Table 5, test of proportions, adjusted $P < 0.001$); most (37 out of 41) were carried by *B. longum* (Fig. 4), and *B. pseudocatenulatum* contained four such gene families (Extended Data Fig. 7, Supplementary Table 5). In samples with *B. longum*, this implicated a clear strain shift after weaning, when fewer *B. longum* strains carried these genes (Fig. 4). In samples with *B. pseudocatenulatum*, four gene families showed a similar but less contrasting pattern (Extended Data Fig. 7). Overall, these observations identify new candidate species that contribute to HMO processing or exploitation, and link strain composition to specific driving molecular functions that potentially explain selective sweeps during microbiome development, in this case specifically related to breastfeeding.

Despite ample sample size, scrutiny of the study design, and thorough statistical analyses, most of the taxonomic and functional signals we detected in case–control comparisons were modest in effect size and statistical significance. This could be due to several reasons—differences between T1D endotypes, temporally diffuse signals, geographical heterogeneity, or lack of stool samples for the first two months of life —and these should be considered in future investigations (Supplementary Note 6). Furthermore, the data used in these investigations was composed of samples from the genetically predisposed and mostly white, non-Hispanic case–control groups designed into the TEDDY study. Results cannot be guaranteed to reflect the whole TEDDY cohort or child populations in the respective countries.

Future targeted approaches to identify subject-specific connections between the gut microbiota and T1D pathogenesis may be beneficial, particularly given the apparent population-level heterogeneity revealed here. For example, laboratory experiments involving dietary factors that have been associated with the onset of T1D[3] may reveal biochemically specific signals that are mediated by the microbiome. Different endotypes of disease, such as differences in the first appearing autoantibody (IAA versus GADA), the number of appearing autoantibodies, the time from seroconversion to T1D diagnosis, genetic host risk alleles and ethnic backgrounds, may be characterized by distinct microbial configurations (Supplementary Note 6). Finally, components of the microbiome that were poorly measured in these data may also have crucial roles: viruses, fungi, microbial transcription or small-molecule biochemistry. By surveying these additional molecular activities by cross-sectional analysis and in more detailed longitudinal populations, this study lays

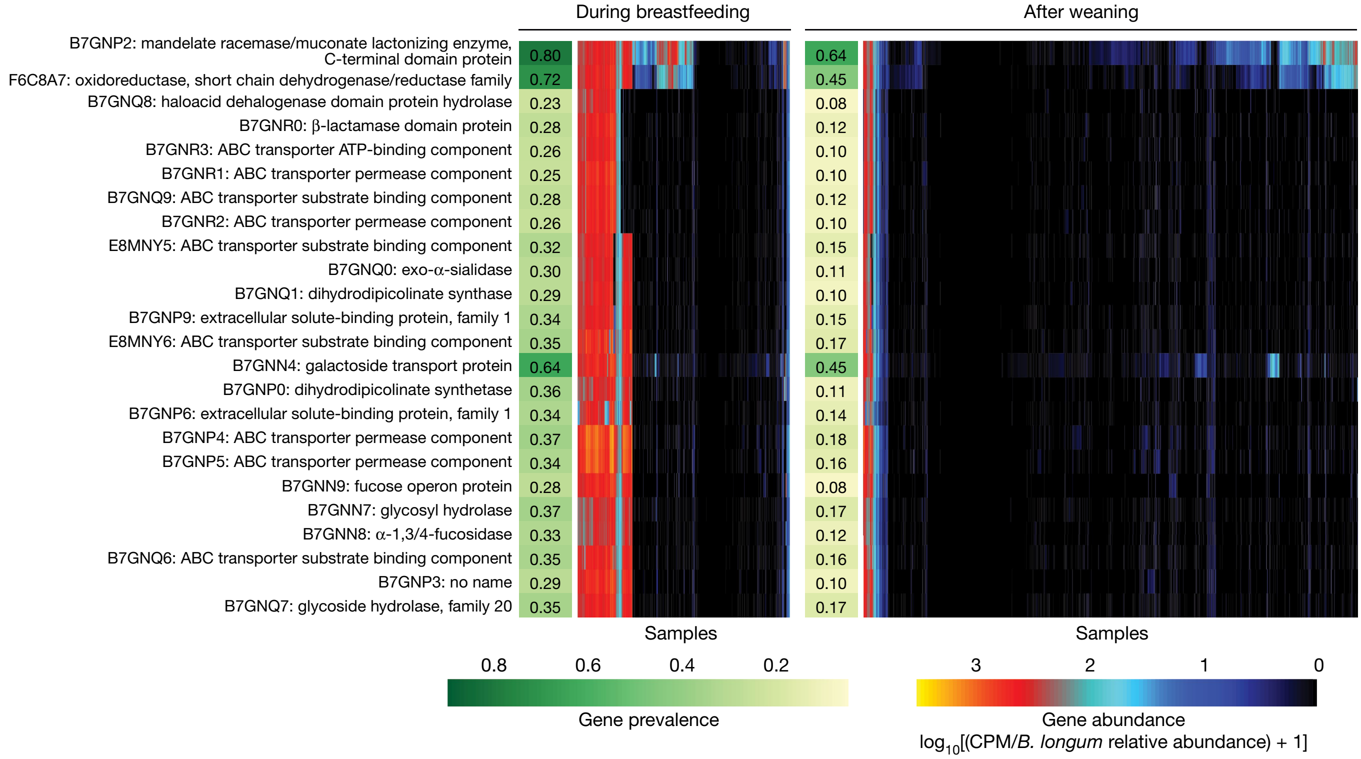


Fig. 4 | *Bifidobacterium longum* strains are characterized by HMO gene content and stratified by breastfeeding status. Gene families involved in HMO utilization and showing contrasting presence in *B. longum* genomes during breastfeeding ($n = 1{,}584$ samples) compared to after weaning ($n = 3{,}705$ samples). Abundance heat map columns represent stool samples in which the relative abundance of *B. longum* species was more than 10% ($n = 5{,}289$ samples). Rows and columns were ordered by hierarchical clustering using the complete linkage method. As in Fig. 3, values reflect units of CPM and were further divided by relative abundance of *B. longum* to obtain quantifications that are comparable between samples. UniRef90 identifiers and gene names or families are indicated on the left.

the foundation to identify further gut microbial components that are predictive, protective or potentially causal in T1D risk or pathogenesis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0620-2.

Received: 16 November 2017; Accepted: 6 September 2018;
Published online 24 October 2018.

1. Katsarou, A. et al. Type 1 diabetes mellitus. *Nat. Rev. Dis. Primers* **3**, 17016 (2017).
2. Pociot, F. & Lernmark, Å. Genetic risk factors for type 1 diabetes. *Lancet* **387**, 2331–2339 (2016).
3. Rewers, M. & Ludvigsson, J. Environmental risk factors for type 1 diabetes. *Lancet* **387**, 2340–2348 (2016).
4. Knip, M. & Siljander, H. The role of the intestinal microbiota in type 1 diabetes mellitus. *Nat. Rev. Endocrinol.* **12**, 154–167 (2016).
5. Hober, D. & Sauter, P. Pathogenesis of type 1 diabetes mellitus: interplay between enterovirus and host. *Nat. Rev. Endocrinol.* **6**, 279–289 (2010).
6. Paun, A., Yau, C. & Danska, J. S. The influence of the microbiome on type 1 diabetes. *J. Immunol.* **198**, 590–595 (2017).
7. de Goffau, M. C. et al. Aberrant gut microbiota composition at the onset of type 1 diabetes in young children. *Diabetologia* **57**, 1569–1577 (2014).
8. de Goffau, M. C. et al. Fecal microbiota composition differs between children with β-cell autoimmunity and those without. *Diabetes* **62**, 1238–1244 (2013).
9. Mariño, E. et al. Gut microbial metabolites limit the frequency of autoimmune T cells and protect against type 1 diabetes. *Nat. Immunol.* **18**, 552–562 (2017).
10. Needell, J. C. & Zipris, D. The role of the intestinal microbiome in type 1 diabetes pathogenesis. *Curr. Diab. Rep.* **16**, 89 (2016).
11. Davis-Richardson, A. G. et al. *Bacteroides dorei* dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Front. Microbiol.* **5**, 678 (2014).
12. Kostic, A. D. et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).
13. Endesfelder, D. et al. Compromised gut microbiota networks in children with anti-islet cell autoimmunity. *Diabetes* **63**, 2006–2014 (2014).
14. Maffeis, C. et al. Association between intestinal permeability and faecal microbiota composition in Italian children with beta cell autoimmunity at risk for type 1 diabetes. *Diabetes Metab. Res. Rev.* **32**, 700–709 (2016).
15. Mejía-León, M. E., Petrosino, J. F., Ajami, N. J., Domínguez-Bello, M. G. & de la Barca, A. M. Fecal microbiota imbalance in Mexican children with type 1 diabetes. *Sci. Rep.* **4**, 3814 (2014).
16. Alkanani, A. K. et al. Alterations in intestinal microbiota correlate with susceptibility to type 1 diabetes. *Diabetes* **64**, 3510–3520 (2015).
17. Soyucen, E. et al. Differences in the gut microbiota of healthy children and those with type 1 diabetes. *Pediatr. Int.* **56**, 336–343 (2014).
18. Endesfelder, D. et al. Towards a functional hypothesis relating anti-islet cell autoimmunity to the dietary impact on microbial communities and butyrate production. *Microbiome* **4**, 17 (2016).
19. Zhao, L. et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science* **359**, 1151–1156 (2018).
20. Markle, J. G. et al. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science* **339**, 1084–1088 (2013).
21. Costa, F. R. et al. Gut microbiota translocation to the pancreatic lymph nodes triggers NOD2 activation and contributes to T1D onset. *J. Exp. Med.* **213**, 1223–1239 (2016).
22. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
23. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, e177 (2007).
24. Koenig, J. E. et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl Acad. Sci. USA* **108**, 4578–4585 (2011).
25. Dominguez-Bello, M. G. et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl Acad. Sci. USA* **107**, 11971–11975 (2010).
26. Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 690–703 (2015).
27. Yassour, M. et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8**, 343ra81 (2016).
28. Hagopian, W. A. et al. The Environmental Determinants of Diabetes in the Young (TEDDY): genetic criteria and international diabetes risk screening of 421 000 infants. *Pediatr. Diabetes* **12**, 733–743 (2011).
29. Lee, H. S. et al. Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Diabetes Metab. Res. Rev.* **30**, 424–434 (2014).
30. Stewart, C. J. et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* https://doi.org/10.1038/s41586-018-0617-x (2018).

31. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
32. Korpela, K. et al. Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children. *Nat. Commun.* **7**, 10410 (2016).
33. Joice, R., Yasuda, K., Shafquat, A., Morgan, X. C. & Huttenhower, C. Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metab.* **20**, 731–741 (2014).
34. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
35. O'Callaghan, A. & van Sinderen, D. Bifidobacteria and their role as members of the human gut microbiota. *Front. Microbiol.* **7**, 925 (2016).
36. Thurston, B., Dawson, K. A. & Strobel, H. J. Pentose utilization by the ruminal bacterium *Ruminococcus albus*. *Appl. Environ. Microbiol.* **60**, 1087–1092 (1994).
37. Uusitalo, U. et al. Association of early exposure of probiotics and islet autoimmunity in the TEDDY Study. *JAMA Pediatr.* **170**, 20–28 (2016).
38. Underwood, M. A., German, J. B., Lebrilla, C. B. & Mills, D. A. *Bifidobacterium longum* subspecies *infantis*: champion colonizer of the infant gut. *Pediatr. Res.* **77**, 229–235 (2015).
39. Sela, D. A. et al. The genome sequence of *Bifidobacterium longum* subsp. *infantis* reveals adaptations for milk utilization within the infant microbiome. *Proc. Natl Acad. Sci. USA* **105**, 18964–18969 (2008).
40. Vatanen, T. et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).

Acknowledgements This research was performed on behalf of the TEDDY Study Group, which is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, UC4 DK100238, UC4 DK106955, UC4 DK112243, UC4 DK117483, and contract no. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. This work supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida (UL1 TR000064) and the University of Colorado (UL1 TR001082). C.H. was supported by funding from JDRF (3-SRA-2016-141-Q-R) and NIDDK (U54DE023798, R24DK110499). H.V. and R.J.X. were supported by funding from JDRF (2-SRA-2016-247-S-B, 2-SRA-2018-548-S-B).

Reviewer information *Nature* thanks K. Aagaard, C. Lozupone and L. Wen for their contribution to the peer review of this work.

Author contributions T.V., E.A.F. and R.S. analysed the metagenomic sequencing data. C.J.S., N.J.A. and J.F.P. generated the metagenomic sequencing data. S.T., T.D.A. and H.V. designed and conducted bacterial growth assays. K.V., Å.L., W.A.H., M.J.R., J.-X.S., J.T., A.-G.Z., B.A. and J.P.K. contributed to the study concept, design and sample acquisition. H.L., H.V., C.H. and R.J.X. served as principal investigators. T.V., E.A.F., H.V., C.H. and R.J.X drafted the manuscript. All authors discussed the results, contributed to critical revisions and approved the final manuscript. Members of the TEDDY Study Group are listed in the Supplementary Information.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0620-2.

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0620-2.

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to T.V. or C.H. or R.J.X.

METHODS

Cohort and study design. TEDDY is a prospective cohort study funded by the National Institutes of Health with the primary goal to identify environmental causes of T1D. It includes six clinical research centres—three in the United States (Colorado, Georgia/Florida, Washington) and three in Europe (Finland, Germany and Sweden). Detailed study design and methods have been previously published[28,41,42]. Written informed consents were obtained for all study participants from a parent or primary caretaker, separately, for genetic screening and participation in a prospective follow-up. The TEDDY study was approved by local US Institutional Review Boards and European Ethics Committee Boards in Colorado's Colorado Multiple Institutional Review Board, Georgia's Medical College of Georgia Human Assurance Committee (2004–2010), Georgia Health Sciences University Human Assurance Committee (2011–2012), Georgia Regents University Institutional Review Board (2013–2015), Augusta University Institutional Review Board (2015–present), Florida's University of Florida Health Center Institutional Review Board, Washington state's Washington State Institutional Review Board (2004–2012) and Western Institutional Review Board (2013–present), Finland's Ethics Committee of the Hospital District of Southwest Finland, Germany's Bayerischen Landesärztekammer (Bavarian Medical Association) Ethics Committee, Sweden's Regional Ethics Board in Lund, Section 2 (2004–2012) and Lund University Committee for Continuing Ethical Review (2013–present). The study is monitored by External Advisory Board formed by the National Institutes of Health.

This analysis used stool samples and clinical metadata from two nested case–control studies (persistent, confirmed IA or T1D) using risk set sampling[29]. The data used here were collected as of 31 May 2012, as a 1:1 match in which one control per case of persistent confirmed IA or T1D was selected from the full TEDDY cohort. A control was a participant who had not developed persistent, confirmed IA or T1D by the time the case to which it was matched had developed IA or T1D, within ±45 days of the event time. Matching factors were clinical centre, sex and family history of T1D to control for differences in geographical area, genetic background and in sample or data handling between clinical centres. In all case–control comparisons, we removed all case–control pairs in which the control later progressed to case status (that is, progressed to IA or T1D). In additional, 17 subjects with missing information about breastfeeding together with their matched pairs were excluded from the case–control comparisons to avoid confounding effects from unknown breastfeeding status.

The development of persistent, confirmed IA was assessed every three months. Persistent autoimmunity was defined by the presence of confirmed islet autoantibody on two or more consecutive visits. The date of persistent autoimmunity was defined as the draw date of the first sample of the two consecutive samples that deemed the child persistently positive for a specific autoantibody (or any autoantibody). T1D was defined according to American Diabetes Association criteria for diagnosis[43].

Stool samples were collected monthly starting at three months of age and continuing up until 48 months of age, then every three months until 10 years of age and then biannually thereafter, into the three plastic stool containers provided by the clinical centre. Children who were antibody negative after 4 years of age were encouraged to submit four times a year even though after 4 years their visits schedule switched to biannual. Parents sent the stool containers at either ambient or +4 °C temperature with guaranteed delivery within 24 h in the appropriate shipping box to the NIDDK repository if living in the United States or their affiliated clinical centre if living in Europe. The European clinical centres stored the stool samples and sent monthly bulk shipments of frozen stool to the NIDDK repository. The TEDDY Manual of Operations, including the stool sample collection protocol, can be accessed online at https://repository.niddk.nih.gov/static/studies/teddy/teddy_moop.pdf.

A priori power calculations using discrete Cox's proportional hazards regression[44] for the matched IA case–control study estimated 80% power, $\alpha = 0.01$, two-sided test to detect an odds ratio > 3 for an exposure with 5% prevalence, to an odds ratio > 1.8 for an exposure with 20% prevalence. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Metagenomic sequencing and initial bioinformatics. Samples were metagenomically sequenced as one library each multiplexed through Illumina HiSeq machines using the 2 × 100-bp paired-end read protocol. Samples with limited DNA quantity and/or too few high-quality reads were filtered out, resulting in a discrepancy of sample frequencies between the metagenomic data and the 16S rRNA amplicon sequencing data analysed in the companion paper[30]. Casava v1.8.2 (Illumina) output initial FASTQ files from the resulting data were processed using cutadapt v1.9dev2 for adaptor removal, Trim Galore v0.2.8 (Babraham Bioinformatics) for removing low-quality bases and PRINSEQ v0.20.3[45] for sample demultiplexing. Bowtie2 v2.2.3 was used to map reads to the human genome for decontamination before subsequent analysis.

Taxonomic and functional profiling by MetaPhlAn2 and HUMAnN2. Taxonomic profiling of the metagenomic samples was performed using MetaPhlAn2[46] v2.6.0, which uses a library of clade-specific markers to provide pan-microbial (bacterial, archaeal, viral and eukaryotic) quantification at the species level. MetaPhlAn2 was run using default settings.

Functional profiling was performed with HUMAnN2[47] v0.9.4. For an input metagenome, HUMAnN2 constructs a sample-specific reference database by concatenating and indexing the pangenomes of species detected in the sample by MetaPhlAn2 (pangenomes are pre-clustered, pre-annotated catalogues of open reading frames found across isolate genomes from a given species[48]). HUMAnN2 then maps sample reads against this database to quantify gene presence and abundance in a species-stratified manner, with unmapped reads further used in a translated search against UniRef90[49] to include taxonomically unclassified but functionally distinct gene family abundances. Finally, for community-total, species-stratified, and unclassified gene family abundance, HUMAnN2 reconstructs metabolic pathway abundance based on the subset of gene families annotated to metabolic reactions (based on reaction and pathway definitions from MetaCyc[50]). Enzyme (level-4 Enzyme Commission (EC) categories) abundances were further computed by summing the abundances of individual gene families annotated to each EC number based on UniRef90-EC annotations from UniProt[51].

Phenotype and covariate analysis. This study includes extensive collection of clinical covariates that cover several aspects of common and rare life events in early childhood from infancy up to five years of age. In these analyses, we used information that is, according to the literature, of high relevance in terms of gut microbiome development. Information about mothers, pregnancy and birth was collected during the three-month clinic visit by questionnaire and included the mode of birth (vaginal birth versus caesarean section), gestational age, infant's 5-min Apgar score, information about maternal diabetes (T1D, T2D or gestational diabetes) and maternal insulin and medication use (antibiotics, angiotensin-converting enzyme inhibitors, metformin, glyburide, antihypertensives) during pregnancy. Dietary information used in these analyses includes the start (and end) date for the following dietary compounds: breastfeeding, baby formula, cow's milk, gluten, cereals, meat, vegetables and fruits. The start of solid food (anything other than breast milk or cow's milk) was also analysed separately. The T1D-associated autoantibodies IAA, GADA and IA2A were analysed from serum samples collected at each clinic visit. In addition to IA, defined as persistent, confirmed autoantibody seropositivity, we analysed the data in terms of the persistency and cumulative frequency of autoantibodies (single or multiple autoantibodies). In TEDDY, all prescribed antibiotic courses are recorded. We further stratified these data by the type of antibiotic in five categories: amoxicillin, penicillin, cephalosporins, macrolide and other antibiotics. Information about probiotics covered the dates for starting and stopping probiotic supplementation, but not the specific types of probiotics used. In addition, sex, information about whether first degree relatives in family had T1D, and HLA haplotypes of the subjects were used in these analyses. Subjects screened from the general population were identified with high-risk alleles (89%) including: DRB1*04-DQA1*03-DQB1*03:02/DRB1*03-DQA1*05-DQB1*02:01 (DR3/4), DRB1*04-DQA1*03-DQB1*03:02/DRB1*04-DQA1*03-DQB1*03:02 (DR4/4), DRB1*04-DQA1*03-DQB1*03:02/DRB1*08-DQA1*04-DQB1*04:02 (DR4/8) and DRB1*03-DQA1*05-DQB1*02:01/DRB1*03-DQA1*05-DQB1*02:01 (DR3/3), plus six genotypes specific to first-degree relatives[28].

Principal coordinate analysis (PCoA) ordination was generated using *t*-distributed stochastic neighbour embedding (*t*-SNE) as implemented in Rtsne package in R with Bray–Curtis dissimilarity as the distance measure and perplexity (a free parameter) equal to 50. Statistical significance of the trends between early clusters and metadata were tested using mixed-effect logistic regression and samples collected during the first year of life as follows. The target variable used was a binary indicator of whether the relative abundance of the taxon of interest (three different *Bifidobacterium* species or phylum Proteobacteria) was greater than 0.5 (definition of the cluster). The age of sample collection, mode of delivery, clinical centre, breastfeeding status (ongoing/stopped), solid food status (binary variable indicating whether solid food was introduced in the diet) and antibiotics status (binary variable indicating whether the subject received antibiotics during the last 30 days) were used as fixed effects, and the subject ID was used as a random effect.

Associations between microbial feature abundances and clinical outcome were determined using MaAsLin[52]. In brief, this multivariate linear modelling system for microbial data selects from among a set of (potentially high-dimensional) covariates to associate with microbial taxon or pathway abundances. Mixed-effects linear models using a variance-stabilizing arcsin square root transform on relative abundances are then used to determine the significance of putative associations from among this reduced set. In the models, subject ID was used as a random effect, and the age of sample collection, mode of delivery, clinical centre (for cohort-wide comparisons), breastfeeding status (ongoing or stopped), solid food status (binary variable indicating whether solid food was introduced in the diet), number of sequencing reads and case–control outcome were used as fixed effects. Nominal

P values were adjusted using the Benjamini–Hochberg FDR method. Here, microbial features with corrected $q < 0.25$ were reported. For metabolic pathways, pseudocount 2^6 was added to CPM values to stabilize the variation in lowly abundant and/or prevalent but highly variable categories, and data were $\log_2$-transformed.

As previously described[40], to associate microbial diversity with covariates while accounting for nonlinear, age-dependent effects, we first fitted a sigmoid function (nls function in R) to account for the longitudinal trend. Residuals of this model were then used as inputs for a mixed-effect model (glmmPQL function in the MASS R package), with subject IDs as random effects to account for repeated measurements in the data. Other factors were included in the model as fixed effects, and their significance levels were evaluated using P values reported by the model (Supplementary Table 2).

The association between T1D case–control outcome and microbial alpha diversity in individual clinical centres was tested using a linear mixed-effects model (glmmPQL function in MASS R package) on samples 730 days or less before T1D diagnosis. In the model, the age at stool sample collection and T1D case–control outcome were used as fixed effects, and subject ID was used as a random effect.

Microbial variance explained by clinical and other covariates. Variance analysis was conducted using the adonis function in the vegan R package given a Bray–Curtis dissimilarity matrix of the taxonomic profiles and all TEDDY clinical metadata listed above. In brief, adonis conducts multivariate ANOVA using the dissimilarity matrix (that is, partitions the sums of squares) given the metadata as covariates. Statistical significance of the fit was assessed using permutation tests.

HMO gene homology. The HMO gene cluster homologues between *B. longum* subsp. *infantis* and multiple taxa were analysed as follows. UniRef90 gene families corresponding to the protein sequences in the *B. longum* subsp. *infantis* HMO gene cluster[39] (protein sequences Blon_2331-Blon_2361 in NCBI protein sequence database) were identified by translated BLAST search against ChocoPhlAn pangenome collection[48] used by HUMAnN2. Identified hits were further filtered by requiring $\geq$50% alignment identity and $\geq$80% mutual coverage. Combining this information with HUMAnN2 species-stratified UniRef90 gene family quantification enabled calling these genes present given that they had sufficient read coverage, here defined as $\log_{10}$(counts per million) > 0.1 in at least 50 samples collected during breastfeeding. Differential gene prevalence during breastfeeding was tested using the samples in which the carrier species had >1% relative abundance. Testing was conducted using the test of equal or given proportions (prop.test function in R) and by comparing the prevalence (proportion of the samples for which the species in question harboured the gene according to the metagenomic data) of the gene in samples collected during breastfeeding with the samples collected after weaning. P values were adjusted for multiple testing by Benjamini–Hochberg method (p.adjust function in R). All homologues together with their BLAST search metrics, prevalence in the metagenomic data and corresponding *B. infantis* HMO gene are reported in Supplementary Table 5.

Bacterial growth assays. *Bifidobacterium bifidum* strain RJX-1201, *Bifidobacterium breve* RJX-1202 and *Bifidobacterium longum* RJX-1203 were streaked on brain heart infusion agar (BD) supplemented with 1% vitamin K/hemin solution (BD; sBHI), and incubated for 48 h in a vinyl anaerobic chamber (Coy Laboratory Products) containing 5% CO_2, 5% H_2 and 90% N_2 and maintained at 37 °C. Cells were transferred to sBHI liquid medium (BHI broth, BD, supplemented as above) and grown for 24 h in anaerobic conditions. Cultures were washed twice with PBS and optical density at 600 nm (OD_{600}) was measured using a BioTek PowerWave 340 plate reader. OD_{600} was normalized to 0.2 for all strains and 5 μl bacteria inoculum was added to a final volume of 200 μl containing 10% sBHI and 125 mM carbon source (glucose, fucose, galactose or lactose) in a 96-well plate. OD_{600} was measured in the plate reader every hour for 48 h with 5 s of medium shaking before each measurement. All of the measurements were normalized to a medium-only blank. Experiment was repeated three times ($n = 3$) in triplicate and one representative experiment is shown. Error bars are s.d. of three technical replicates.

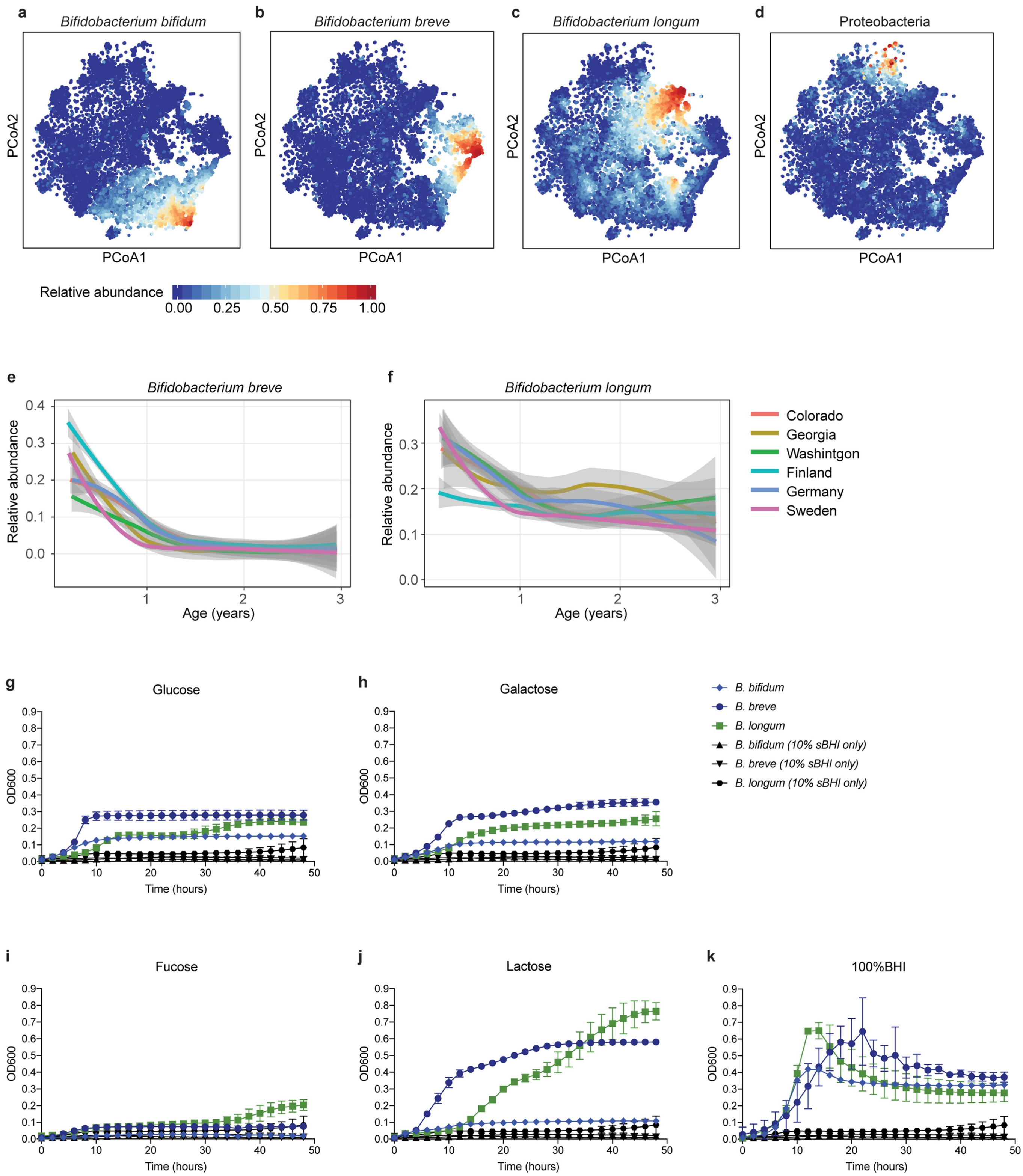
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Code for Random Forest case–control comparisons and cohort wide MaAsLin association analyses in Supplementary Table 4 has been made publicly available at https://github.com/tvatanen/broad_teddy_microbiome_analyses. Other analysis software including quality control, taxonomic, and functional profilers is publicly available and referenced as appropriate.

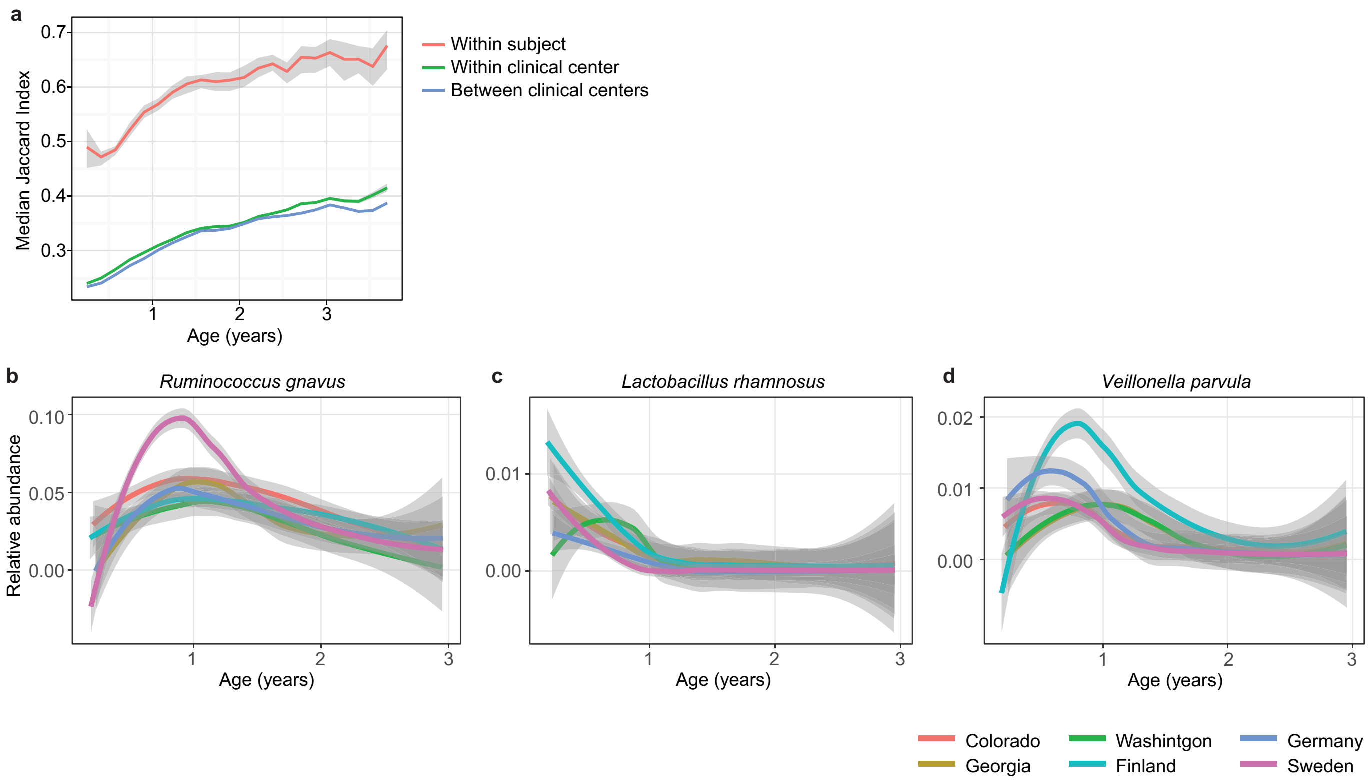
Data availability

TEDDY microbiome 16S and whole-genome sequencing data that support the findings of this study are available in the NCBI database of Genotypes and Phenotypes (dbGaP) with the primary accession code phs001443.v1.p1, in accordance with the dbGaP controlled-access authorization process. Clinical metadata analysed during the current study are available in the NIDDK Central Repository at https://www.niddkrepository.org/studies/teddy.

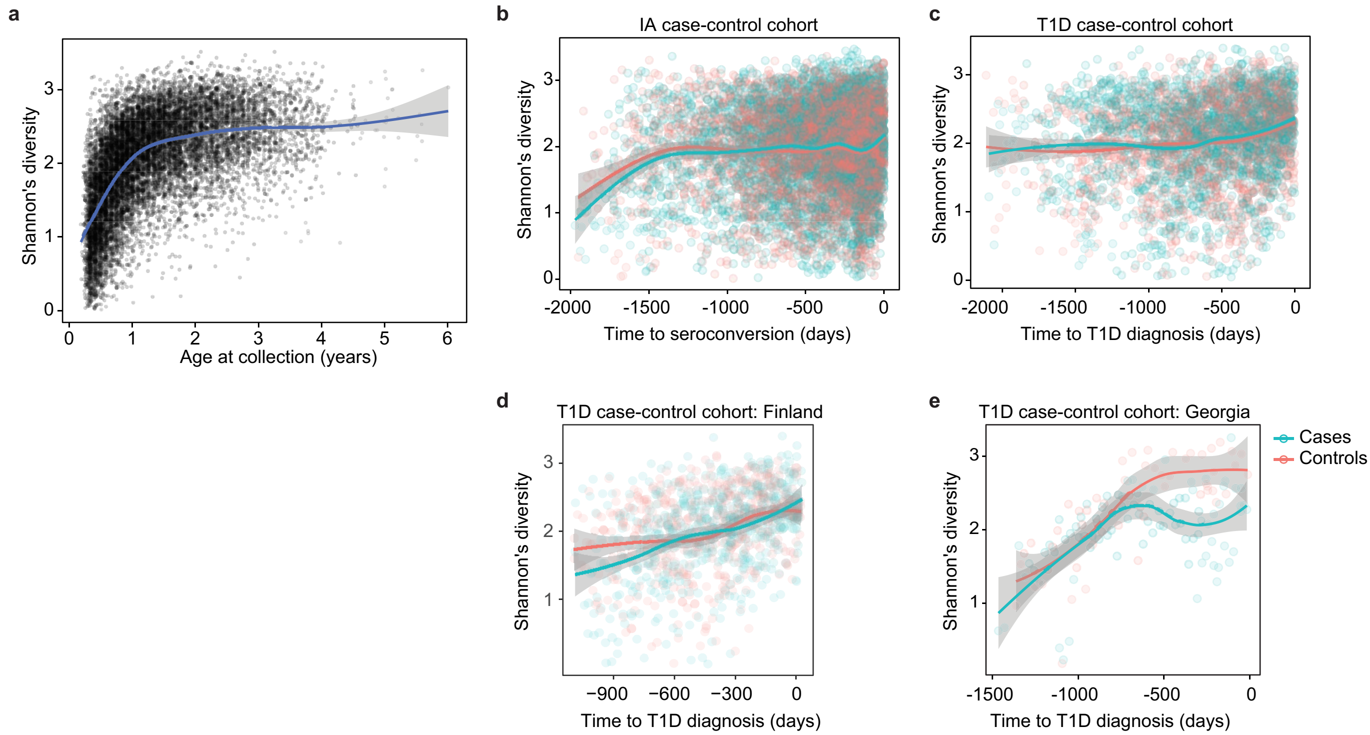
41. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) study: study design. *Pediatr. Diabetes* **8**, 286–298 (2007).
42. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) Study. *Ann. NY Acad. Sci.* **1150**, 1–13 (2008).
43. American Diabetes Association. 2. Classification and diagnosis of diabetes. *Diabetes Care* **38**, S8–S16 (2015).
44. Lachin, J. M. Sample size evaluation for a multiply matched case–control study using the score test from a conditional logistic (discrete Cox PH) regression model. *Statist. Med.* **27**, 2509–2534 (2012).
45. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
46. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
47. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLOS Comput. Biol.* **8**, e1002358 (2012).
48. Huang, K. et al. MetaRef: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids Res.* **42**, D617–D624 (2014).
49. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
50. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).
51. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
52. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).



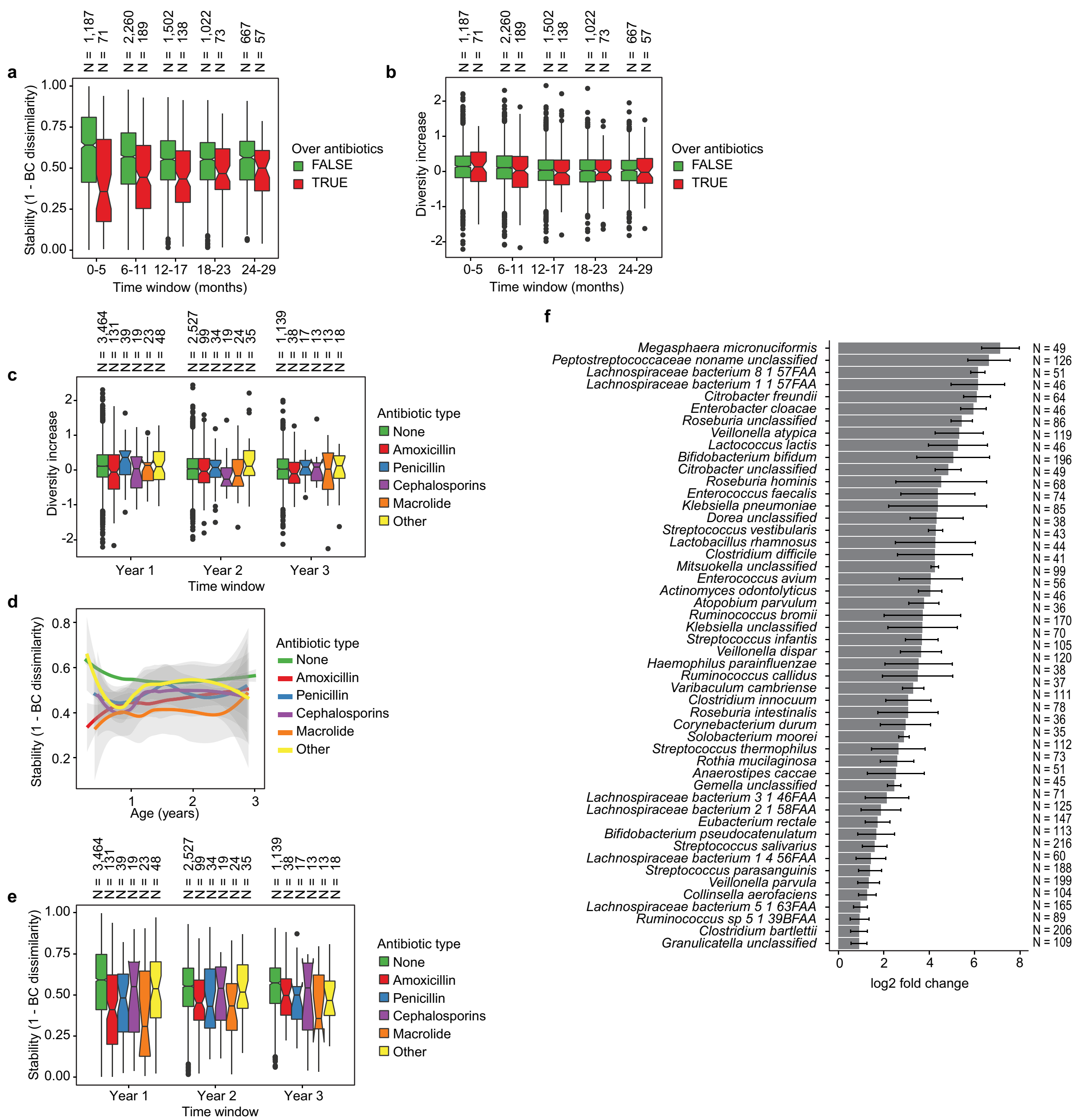
Extended Data Fig. 1 | Heterogeneity in early taxonomic profiles. **a–d**, Relative abundances of taxonomic groups highlighted by weighted averages in Fig. 2a (arrows) shown separately ($n = 10{,}913$ samples). **e**, **f**, Average longitudinal abundance of *B. breve* (**e**) and *B. longum* (**f**) per clinical centre ($n = 10{,}194$ samples). The curves show LOESS fits for the relative abundances, and shaded area shows 95% confidence interval for each fit, as implemented in geom_smooth function in ggplot2 R package. **g–k**, Growth curves of human infant isolates of *B. breve*, *B. bifidum* and *B. longum* grown individually in low-nutrient medium (10% sBHI) supplemented with single carbon sources (glucose (**g**), galactose (**h**), fucose (**i**) and lactose (**j**)) or grown in 100% sBHI (**k**). As a negative control, growth curves of each strain grown in 10% BHI without additional sugar are shown in black for each condition. Data are representative of three independent experiments and are presented as the mean and s.d. of triplicate assessments.



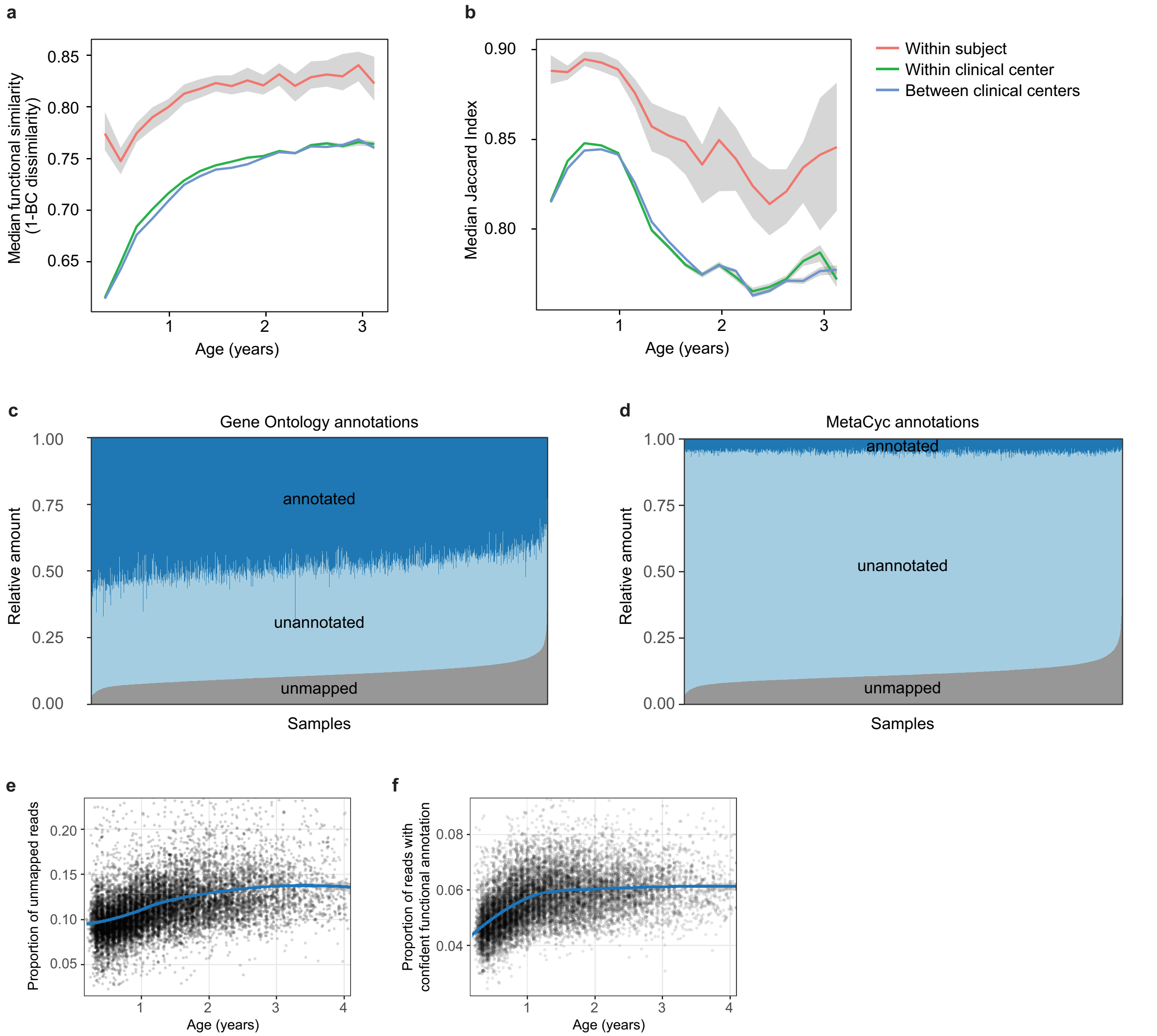
Extended Data Fig. 2 | Stability and regional differences of taxonomic profiles. a, Stability of the microbiota, measured by the Jaccard index ($n = 10{,}750$ samples) in three-month time windows, over two-month increments, stratified into three groups: within subject, within clinical centre, and across clinical centres. Lines show the median per time window. Shaded areas show the 99% confidence interval estimated using binomial distribution. Compare to Fig. 2b, which shows the same analysis measured by Bray–Curtis dissimilarity. **b**–**d**, Average longitudinal abundance of *Ruminococcus gnavus* (**b**), *Lactobacillus rhamnosus* (**c**) and *Veillonella parvula* (**d**) per clinical centre ($n = 10{,}194$ samples). The curves show LOESS fit for the relative abundances, as above.



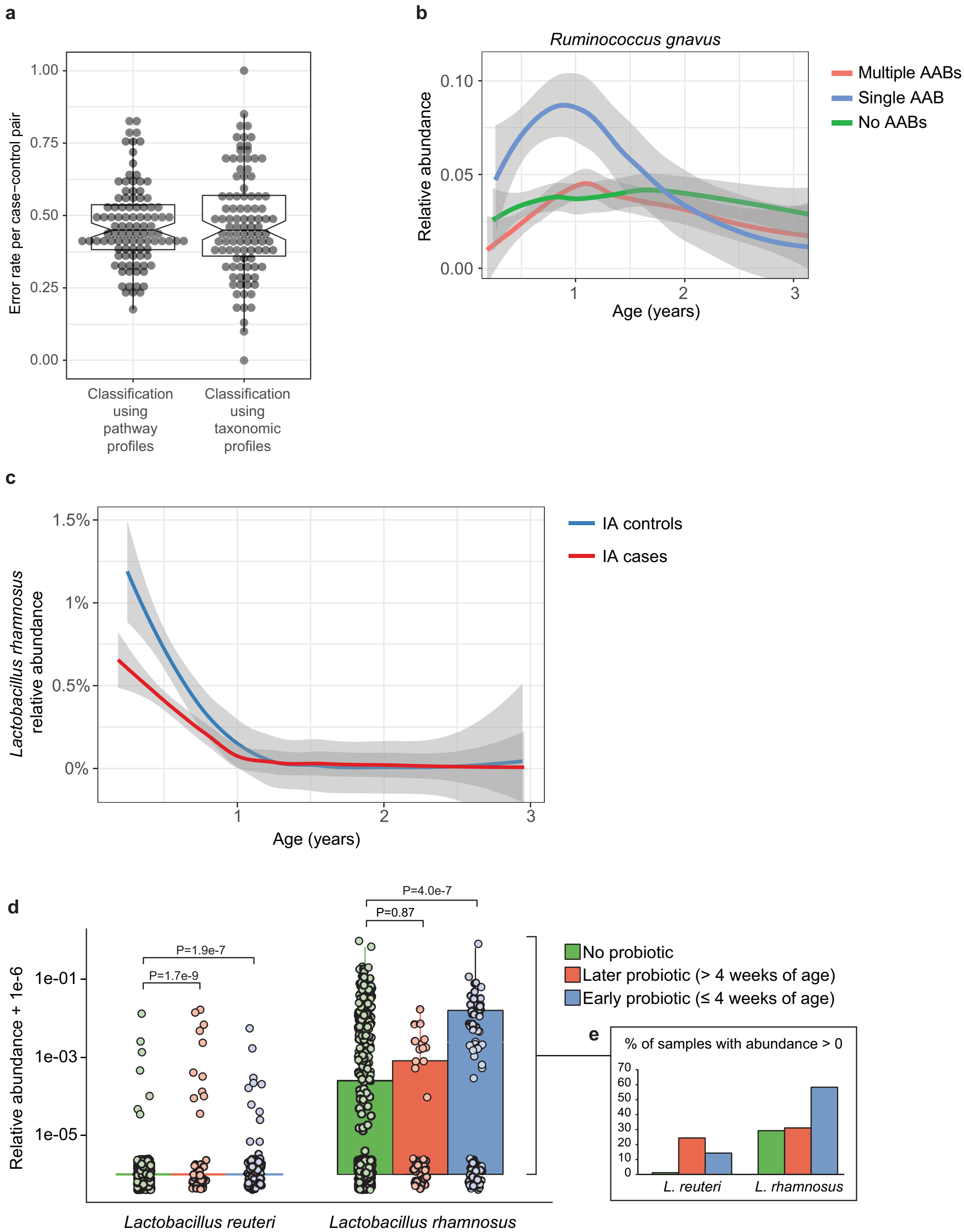
Extended Data Fig. 3 | Accrual of microbial alpha diversity. a, Shannon's diversity of the taxonomic profiles of the gut microbial communities ($n = 10{,}913$ samples) with respect to the age at the sample collection. The curve shows the generalized additive model (GAM) fit for the data, and the shaded area shows the 95% confidence interval for each fit, as implemented in geom_smooth function in ggplot2 R package. **b**, Shannon's diversity for the samples in the IA case–control cohort ($n = 7{,}051$) with respect to the time to the appearance of first autoantibody (seroconversion). The curves show LOESS fits for cases and controls separately, and the shaded area shows 95% confidence intervals for each fit. **c**, Shannon's diversity for the samples in the T1D case–control cohort ($n = 3{,}309$) with respect to the time to T1D diagnosis. The curves and shaded areas are as in **b**. **d**, As in **c**, but only for data ($n = 983$ samples) for subjects in Finland. No difference between cases and controls. **e**, As in **c**, but only for data ($n = 142$ samples, $n = 6$ subjects) for subjects in Georgia, USA. Cases show a drop in alpha diversity before the diagnosis of T1D (linear mixed-effects model, $P = 0.0033$).



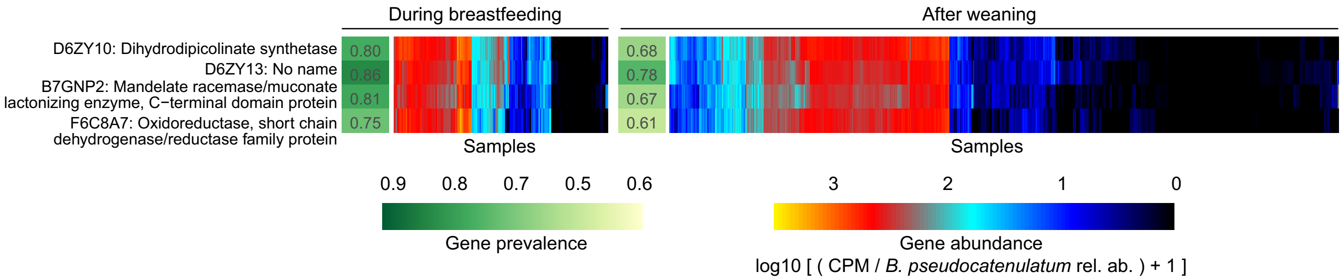
Extended Data Fig. 4 | Effects of antibiotics. **a**, Influence of antibiotic courses on microbial stability, stratified into six-month time windows (x axis). Stability was measured by Bray–Curtis dissimilarity over consecutive stool samples (<50 days apart) from the same individual between 3 and 29 months of age, and stratified by whether antibiotics were given between the two samples. For each notched box plot, the box denote the interquartile range (IQR), the horizontal line denotes the median, and the notch denotes the approximation for the 95% confidence interval (notch width $= 1.58 \times \mathrm{IQR}/n^{0.5}$, in which n is the number of samples per box plot). Compare to Fig. 2c. **b**, Influence of antibiotic courses on microbial diversity. Notched box plots denote the increase (difference) in diversity between two consecutive stool samples (<50 days apart) stratified by antibiotic administration between the samples. Data show no difference between the groups (antibiotics versus no antibiotics). **c**, Influence of antibiotics courses on microbial diversity by antibiotic type; data from **b** stratified into one-year time windows (x axis) and antibiotic types. No significant differences were detected between the antibiotic types. **d**, **e**, Influence of antibiotic courses on microbial stability by antibiotic type; data from Fig. 2c and Extended Data Fig. 3a stratified by antibiotic type. **d**, LOESS fit for the relative abundances (shaded area shows 95% confidence interval for each fit, as implemented in geom_smooth function in ggplot2 R package). **e**, Notched box plots (as in **a** and **b**) for the data per antibiotic type. No significant differences were detected between the antibiotic types. No antibiotics, $n = 7{,}130$; amoxicillin, $n = 268$; penicillin, $n = 90$; cephalosporin, $n = 51$; macrolide, $n = 60$; other, $n = 101$. **f**, Decreases in relative abundance of bacteria over antibiotic courses. Bacteria for which the bootstrapped 95% confidence interval of the fold change does not overlap zero are shown. Fold change was measured between consecutive samples with an antibiotic course between them, given that the species in question was present in the first of the two samples. Sample size per species (n) indicate the number of sample pairs in which the species in question was present in the sample preceding the antibiotic treatment. Bars denote bootstrapped mean $\log_2$(fold change) (that is, decrease), and error bars denote s.d. ($n = 1{,}000$ bootstrap samples).



Extended Data Fig. 5 | Dynamics of species-specific microbial functional potential during early gut development. a, **b**, Stability of microbial pathways ($n = 10{,}580$ samples) measured by Bray–Curtis dissimilarity (**a**) and the Jaccard index (**b**) and stratified into three groups: within subject, within clinical centre, and across clinical centres. Although the baseline level of functional similarity is significantly greater than that of taxa (see Fig. 2b), functional states and development trajectories also both retain a level of personalization. The stability of the functional profiles was evaluated in three-month time windows, over two-month increments. Lines show the median per time window, and shaded area denotes the 99% confidence interval estimated using binomial distribution. **c**, **d**, Proportion of metagenomic gene abundance with functional annotation through Gene Ontology (**c**) and MetaCyc (**d**) databases. The metagenomic reads were divided into the following categories: reads that could be mapped to genes with functional assignment in the database in question (annotated), and reads with no annotation but alignment to species pangenomes or UniProt proteins (unannotated). The proportion of the unknown genes (unmapped) was estimated using the number of reads with unknown origin. **e**, The proportion of unmapped reads, reflecting the relative abundances of reads not mappable to any microbial pangenomes in the available reference set or to UniProt. An increasing trend of unmapped reads with respect to the age at sample collection continued through approximately two years of age. **f**, The proportion of reads with confident functional annotation in MetaCyc within the genes that mapped to species pangenomes or UniProt proteins. The data again showed an increasing longitudinal trend, implicating a deficit of functional and biochemical annotations within microorganisms that are abundant during the first year of life.



Extended Data Fig. 6 | Differences between cases and controls. a, The gut microbiome functional (left) and taxonomic (right) profiles were classified between cases and controls using leave-one-out cross-validation ($n = 3,366$ samples), in which one case–control pair was held-out in turn. Data show error rates for classifying these held-out samples per fold (a data point per fold, $n = 100$ folds). This suggests weak but better-than-random classification between cases and controls. Notched box plots are as in Extended Data Fig. 4. **b**, Average longitudinal abundance of *Ruminococcus gnavus* in Finland ($n = 2,630$ samples) stratified by the number of observed persistent autoantibodies (AABs); no autoantibodies (that is, healthy control), a single autoantibody, or multiple (two or more) autoantibodies. **c**, Average longitudinal abundance of *Lactobacillus rhamnosus* in IA cases and controls ($n = 7,017$ samples). *L. rhamnosus* is more abundant in controls ($q = 0.055$). The curves in **b** and **c** show LOESS fit per group, and shaded areas show 95% confidence interval for each fit, as implemented in geom_smooth function in ggplot2 R package. **d**, Abundance (left) and prevalence (right) of *Lactobacillus reuteri* and *L. rhamnosus* in the first stool sample of each individual (collected at approximately three months of age) in association with early probiotic supplementation. 'No probiotic' indicates no probiotics given before the first stool sample ($n = 583$); 'later probiotic' refers to probiotics given later than the first four weeks but before the first stool sample ($n = 45$); 'early probiotic' refers to probiotics given during the first four weeks of life ($n = 84$). Numbers (n) per clinical centre are given in Extended Data Table 2. *L. reuteri* and *L. rhamnosus* were more abundant and prevalent in groups with probiotics supplementation. Visual jitter was added to make data equal to zero distinguishable, and boxes denote the IQR, when applicable. The shown P values were obtained by applying Fisher's exact test (two-sided) to presence or absence count data (counting samples in which the species were present).



Extended Data Fig. 7 | Contrasting HMO utilization genes in *B. pseudocatenulatum*. The gene families involved in HMO utilization and that show contrasting presence in *B. pseudocatenulatum* genomes during breastfeeding ($n = 321$ samples) compared to after weaning ($n = 1{,}004$ samples). Columns represent stool samples in which the relative abundance of *B. pseudocatenulatum* species was greater than 10% ($n = 1{,}325$ samples). Rows and columns were ordered by hierarchical clustering using complete linkage method. Compare to Fig. 4, which shows similar data for *B. longum*. UniRef90 identifiers and gene names or families are indicated on the left.

Extended Data Table 1 | Summary of TEDDY microbiome cohort

	US, Colorado	US, Georgia	US, Washington	Finland	Germany	Sweden
T1D cases (samples)	14 (274)	3 (89)	8 (111)	34 (553)	13 (246)	29 (532)
IA cases (samples)	39 (689)	17 (252)	25 (368)	70 (900)	21 (292)	95 (1,542)
Healthy controls (samples)	61 (906)	22 (250)	36 (399)	119 (1,273)	40 (512)	137 (1,725)
Sex						
Male / Female	61 / 53	19 / 23	51 / 18	117 / 106	30 / 44	152 / 109
Ethnic background						
White, non-hispanic	86 (75.4%)	41 (97.6%)	56 (81.2%)	N/A	N/A	N/A
Mode of birth						
Caesarean section	41 (36.0%)	22 (52.4%)	25 (36.2%)	42 (18.8%)	23 (31.1%)	46 (17.6%)
Probiotic supplementation						
Probiotics during first 4 weeks	0	2 (4.8%)	0	67 (30.0%)	7 (9.5%)	14 (5.4%)
Probiotics during follow-up	22 (19.3%)	13 (31.0%)	9 (13.0%)	162 (72.6%)	33 (44.6%)	58 (22.2%)
Breastfeeding						
Median duration (days)	268	301	335	289	278	228
duration, 25 percentile	56	145	171	152	140	98
duration, 75 percentile	396	365	440	385	367	304
Number of subjects never breastfed	3	3	1	0	0	0
Maternal characteristics						
Maternal T1D	7 (6.1%)	0	3 (4.3%)	14 (6.3%)	18 (24.3%)	7 (2.7%)
Maternal T2D	2 (1.8%)	0	0	0	0	0
Gestational diabetes	5 (4.4%)	5 (11.9%)	5 (7.2%)	32 (14.3%)	3 (4.1%)	6 (2.3%)
Antibiotics during pregnancy	21 (18.4%)	10 (23.8%)	5 (7.2%)	40 (17.9%)	13 (17.6%)	29 (11.1%)
Metformin during pregnancy	1 (0.9%)	0	0	1 (0.4%)	0	0
Glyburide during pregnancy	2 (1.8%)	2 (4.8%)	2 (2.9%)	0	0	0
Antihypertensives during pregnancy	4 (3.5%)	3 (7.1%)	4 (5.8%)	5 (2.2%)	3 (4.1%)	0
Insulin during pregnancy	9 (7.9%)	0	3 (4.3%)	23 (10.3%)	19 (25.7%)	8 (3.1%)

Data on subjects' ethnic background were not systematically collected in European clinical centres but these study populations were predominantly white, non-Hispanic. Reported antihypertensive drugs were atenolol ($n=2$), bisoprolol ($n=1$), labetalol ($n=6$), methyldopa ($n=1$), methyldopa plus methyldopate ($n=3$), metoprolol ($n=4$) and nifedipine ($n=5$). No use of angiotensin-converting enzyme (ACE) inhibitors was reported. Numbers indicate the number of subjects (n) if not specified otherwise.

Extended Data Table 2 | Antibiotics and probiotics

	US, Colorado	US, Georgia	US, Washington	Finland	Germany	Sweden
Subjects with abx prescriptions	93 (81.6%)	37 (88.1%)	54 (78.3%)	206 (92.4%)	56 (75.7%)	192 (73.6%)
Median number of abx per subject (25th and 75th percentile)	2 (1-6)	5 (2-9)	2 (1-4)	6 (3-11)	2 (0-5)	2 (0-4)
Number of abx by type (prescriptions per subject)						
Amoxicillin	242 (2.12)	147 (3.50)	104 (1.51)	769 (3.45)	45 (0.61)	134 (0.51)
Cephalosporins	87 (0.76)	65 (1.55)	31 (0.45)	127 (0.57)	51 (0.69)	23 (0.09)
Macrolide	54 (0.47)	35 (0.83)	47 (0.68)	203 (0.91)	33 (0.45)	23 (0.09)
Penicillin	6 (0.05)	2 (0.05)	3 (0.04)	17 (0.08)	13 (0.18)	412 (1.58)
Other	76 (0.67)	80 (1.90)	33 (0.48)	521 (2.34)	77 (1.04)	154 (0.59)
Total	465 (4.08)	329 (7.83)	218 (3.16)	1,637 (7.34)	219 (2.96)	746 (2.86)
Probiotic use in early life						
Early probiotic	0 (0.0%)	1 (2.9%)	0 (0.0%)	63 (30.7%)	7 (10.0%)	13 (5.6%)
Later probiotic	1 (0.9%)	1 (2.9%)	2 (3.3%)	16 (7.8%)	8 (11.4%)	17 (7.3%)
No probiotic	109 (99.1%)	32 (94.1%)	59 (96.7%)	126 (61.5%)	55 (78.6%)	202 (87.1%)

Top, 3,678 antibiotic prescriptions in the TEDDY microbiome study population by clinical centre. Bottom, early probiotic supplementation in TEDDY clinical centres. Probiotic use was stratified into three categories: probiotics during the first 4 weeks of life (early probiotic); probiotics before the first stool sample (roughly at three months) but not the first 4 weeks (later probiotic); and no probiotics before the first stool sample (no probiotic). Data for probiotics are presented as *n* (percentage).

https://doi.org/10.1038/s41586-018-0581-5

OTX2 restricts entry to the mouse germline

Jingchao Zhang[1,4], Man Zhang[1,4]*, Dario Acampora[2], Matúš Vojtek[1], Detian Yuan[3], Antonio Simeone[2] & Ian Chambers[1]*

The successful segregation of germ cells from somatic lineages is vital for sexual reproduction and species survival. In the mouse, primordial germ cells (PGCs), precursors of all germ cells, are induced from the post-implantation epiblast[1]. Induction requires BMP4 signalling to prospective PGCs[2] and the intrinsic action of PGC transcription factors[3–6]. However, the molecular mechanisms that connect BMP4 to induction of the PGC transcription factors that are responsible for segregating PGCs from somatic lineages are unknown. Here we show that the transcription factor OTX2 is a key regulator of these processes. Downregulation of *Otx2* precedes the initiation of the PGC programme both in vitro and in vivo. Deletion of *Otx2* in vitro markedly increases the efficiency of PGC-like cell differentiation and prolongs the period of PGC competence. In the absence of *Otx2* activity, differentiation of PGC-like cells becomes independent of the otherwise essential cytokine signals, with germline entry initiating even in the absence of the PGC transcription factor BLIMP1. Deletion of *Otx2* in vivo increases PGC numbers. These data demonstrate that OTX2 functions repressively upstream of PGC transcription factors, acting as a roadblock to limit entry of epiblast cells to the germline to a small window in space and time, thereby ensuring correct numerical segregation of germline cells from the soma.

Different species form their germ cells by either of two general methods: segregation of preformed germplasm, or induction by signalling[7,8]. In mammals, germ cell precursors arise by induction[9–11]. In the mouse, competence to initiate germ cell development is restricted to a few cells within the E5.5–E6.25 epiblast[1]. BMP4 from the extraembryonic ectoderm acts on these competent cells to specify germ cell identity[2]. Specification also requires transcription factors, notably BLIMP1, AP2γ and PRDM14[3–6]. However, the molecular mechanisms that connect the exposure of competent cells to BMP4 to the activation of PGC transcription factors are obscured by limited access to the peri-implantation embryo. Recently, a system for differentiation of PGC-like cells (PGCLCs) from embryonic stem cells (ESCs) via germline competent epiblast-like cells (EpiLCs)[12] has opened up investigation of molecular events segregating germline and soma.

During the ESC to EpiLC transition, the transcription factor OTX2 becomes expressed and redirects binding of OCT4 to genomic regulatory elements[13,14]. OTX2 was previously characterized as a regulator of anterior patterning[15,16]. Previous work has demonstrated that OTX2 and NANOG have antagonistic functions in ESCs[17,18]. A positive role for NANOG in PGCLC differentiation has now also been added to the known requirements for *Blimp1*, *Prdm14* and *Tfap2c* (also known as *Ap2γ*)[19–21]. We therefore assessed expression of the corresponding mRNAs following addition of PGCLC-inducing cytokines to EpiLCs (Fig. 1a, b). *Blimp1*, *Prdm14* and *AP2γ* mRNAs did not change during the first 12 h. A modest increase in *Ap2γ* mRNA at 24 h preceded more pronounced increases in all three mRNAs by 48 h (Fig. 1b). By contrast, *Otx2* mRNA dropped to around 20% of the EpiLC level at 24 h (Fig. 1b). Immunofluorescence analysis indicated that the proportion of cells expressing OTX2 protein decreased at 24 h, with almost no OTX2-expressing cells detected at 48 h (Fig. 1c, Extended Data Fig. 2a, b). Cultures in which PGCLC cytokines were omitted lost OTX2-expressing cells more slowly (Extended Data Fig. 2a, b). Moreover, while *Otx2* mRNA declines upon FGF/Activin withdrawal, the kinetics of suppression are enhanced by PGCLC cytokine addition (Extended Data Fig. 2d). This suggests that PGCLC cytokines directly repress *Otx2* transcription, a notion supported by the prompt decline in *Otx2* pre-mRNA upon switching EpiLCs into PGCLC media (Extended Data Fig. 2e). BLIMP1 and AP2γ proteins were initially detectable at 24 h, but only in cultures treated with cytokines (Extended Data Fig. 2a, b) and only in cells with reduced OTX2 (Fig. 1c, d, Extended Data Fig. 2c). These results suggest that before the PGC gene regulatory network (GRN) becomes activated, the transcriptional circuitry of the formative pluripotent[22], germline competent[23] state characterized by OTX2 expression[13] becomes extinguished.

The reciprocal relationship of OTX2 with BLIMP1 and AP2γ during PGCLC induction prompted determination of whether a similar spatio-temporal relationship between changes in expression of OTX2 and PGC transcription factors held in vivo. Whole-mount immunofluorescence of pre-streak stage embryos indicated that all epiblast cells express both OCT4 and OTX2 (Fig. 1e). At early- to mid-streak, OTX2 remains widely expressed in the epiblast except for cells showing incipient FRAGILIS[24] and BLIMP1[3] expression (Fig. 1f, h). By late-streak- to early-bud-stage, BLIMP1 is clearly detectable within the FRAGILIS field in cells lacking OTX2 (Fig. 1g, i). These results indicate that OTX2 is expressed ubiquitously in pre-streak epiblast cells but is specifically downregulated in the prospective PGC population before BLIMP1 expression.

Cytokine addition is required for PGCLC differentiation[12]. To assess when, cells were treated for one, two or six days with cytokines and analysed by FACS for surface expression of SSEA1 and CD61, which together act as a marker for PGCLCs[12]. Cytokine treatment for the first day induces around half of the potentially responsive population to express both CD61 and SSEA1 (Extended Data Fig. 2f). Two or six days of cytokine treatment were equally effective at inducing CD61 and SSEA1 (Extended Data Fig. 2f). This suggests that cytokine exposure reaches maximum efficacy at two days, the time required to reduce OTX2 to minimal levels and initiate PGCLC transcription factor expression (Fig. 1b, d, Extended Data Fig. 2a–c).

To directly assess whether OTX2 downregulation influences entry of pluripotent cells into the germline, *Otx2*-null cells were examined. A transgenic Oct4ΔPE::GFP reporter activated upon germline entry[25] was added to *Otx2*$^{fl/-}$ and *Otx2*$^{-/-}$ ESCs[17] (Extended Data Fig. 1). Compared to *Otx2* heterozygotes, *Otx2*$^{-/-}$ cell populations showed widespread activation of Oct4ΔPE, with essentially all cells activating GFP (Fig. 2a). Furthermore, the SSEA1^{+}CD61^{+} cell number is increased 5–10-fold in *Otx2*$^{-/-}$ versus *Otx2*$^{+/+}$ cells (Fig. 2b). This was also the case in independently generated *Otx2*$^{-/-}$ ESCs (Extended Data Fig. 3b) and in additional independent *Otx2*$^{-/-}$ ESCs generated using CRISPR/Cas9 (Extended Data Fig. 1, 3c). Three new ESC lines lacking OTX2 protein (Extended Data Fig. 3d) showed enhanced CD61 and SSEA1 expression during PGCLC differentiation (Extended Data Fig. 3e). These results confirm that a lack of OTX2 promotes germline differentiation.

To investigate at which stage of differentiation OTX2 influences germline entry, an Otx2-ER^{T2} transgene (enabling tamoxifen-induced

[1]MRC Centre for Regenerative Medicine, Institute for Stem Cell Research, School of Biological Sciences, University of Edinburgh, 5 Little France Drive, Edinburgh, EH16 4UU, Scotland. [2]Institute of Genetics and Biophysics "Adriano Buzzati-Traverso", CNR, Via P. Castellino, 111, 80131, Naples, Italy. [3]Department of Biochemistry and Molecular Biology, Shandong University School of Medicine, Jinan, 250012, PR China. [4]These authors contributed equally: Jingchao Zhang, Man Zhang. *e-mail: ichambers@ed.ac.uk; mzhang33@ed.ac.uk

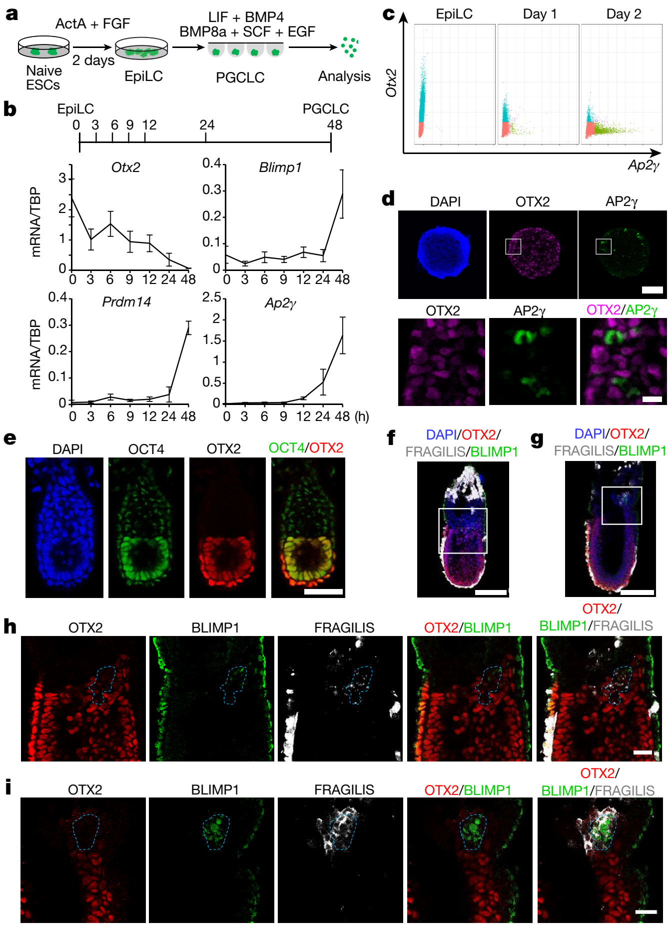


Fig. 1 | Otx2 expression is downregulated before expression of PGC transcription factors. a, Scheme for PGCLC differentiation. **b**, Top, scheme illustrating the time-points (hours) during PGCLC differentiation when mRNAs were analysed. Bottom, real-time PCR (RT-PCR) of *Otx2* and PGC transcription factors in wild-type E14Tg2a ESCs. Expression levels are normalized to TBP; values are means ± s.d.; $n = 3$ biologically independent replicates. **c**, Single-cell quantification of immunofluorescence for OTX2 and AP2γ in cytospin preparations of EpiLCs and cell aggregates at day 1 and day 2 of PGCLC induction. Two biologically independent replicates were performed. **d**, Whole-mount immunofluorescence of E14Tg2a aggregates after one day of PGCLC differentiation; $n = 3$; scale bars, 50 μm (top), 10 μm (bottom). **e**–**g**. Representative confocal images of whole mount staining of embryos at pre-streak (**e**, $n = 4$), early streak (**f**, $n = 3$) and late streak (**g**, $n = 3$) stages; scale bars, 40 μm (**e**), 100 μm (**f**, **g**). **h**, **i**, Magnified image of the regions highlighted in **f** and **g**, respectively. OTX2-negative cells expressing BLIMP1 and FRAGILIS are outlined (**g**, **h**); scale bar, 20 μm.

re-localization of OTX2) was introduced into *Otx2*$^{lacZ/GFP}$ ESCs (Fig. 2c, Extended Data Fig. 4a). Tamoxifen treatment for the first two days suppressed emergence of SSEA1$^+$CD61$^+$ cells (Fig. 2d). These results establish that enforcing OTX2 activity at a time when cells are competent to enter the germline[12] and when cytokines act to decrease endogenous *Otx2* expression, is sufficient to block cytokine-mediated PGCLC differentiation.

To examine the mechanism by which OTX2 inhibits PGCLC differentiation, mRNAs were analysed. Expression of PGCLC transcription factor mRNAs occurred sooner and to a greater extent without OTX2 (Fig. 2e). In contrast, enforcing OTX2 activity inhibited induction of PGCLC transcription factor mRNAs (Fig. 2e). NANOG also directs PGCLC differentiation[20,26]. Consistent with this, endogenous *Nanog* mRNA was induced precociously in *Otx2*$^{-/-}$ cells (Fig. 2e). In addition, although *Fgf5*, *Foxd3* and *Oct6* mRNAs were not induced without OTX2 (Extended Data Fig. 3f), their expression was increased above wild-type levels by OTX2 induction (Extended Data Fig. 3f). These analyses suggest that OTX2 acts at the juncture between somatic and germline differentiation and inhibits PGCLC differentiation by preventing PGC transcription expression.

A previous report has provided evidence for the involvement of T (Brachyury) in PGCLC induction[27] by showing that BMP4 induced T expression via endogenous Wnt. We also found that T is activated robustly only when BMP4 is present (Extended Data Fig. 4b) with T and other somatic markers induced during PGCLC differentiation (Extended Data Fig. 4c). In vivo, BMP4 induces epiblast cells to secrete Wnt[28]. Therefore, to assess whether Wnt acts as an intermediary between BMP4 and activation of T and other somatic markers, Wnt signalling was mimicked by adding CHIR99021 to basal media (Extended Data Fig. 4d). *T*, *Hoxa1* and *Hoxb1* mRNAs were induced by CHIR99021 but *Otx2* mRNA was repressed; effects that were reversed by addition of the Wnt antagonist XAV939 (Extended data Fig. 4d). Therefore, Otx2 downregulation by BMP4 may occur via Wnt signalling. To further assess this, XAV939 was added during PGCLC differentiation. XAV939 did not affect *Otx2* mRNA for 9 h, but dampened further reduction (Extended Data Fig. 4e). Moreover, XAV939 diminished induction of *Blimp1* and *Prdm14* mRNAs seen after 24 h (Extended Data Fig. 4e). This is consistent with a model in which BMP induction of Wnt enforces timely repression of *Otx2* and full induction of *Blimp1* and *Prdm14*. Finally, *Otx2*$^{-/-}$ cells did not activate T mRNA (Extended Data Fig. 4c) or protein (Extended Data Fig. 4f). Therefore, activation of T is dispensable for PGC induction, at least when OTX2 is absent. These observations suggest that the effects of Wnt signalling during PGC differentiation could be attributed to OTX2 downregulation.

To assess whether OTX2 can interfere with the function of an established PGCLC GRN, OTX2 activity was restored at day 2, once the PGC GRN was already activated (Fig. 2e). This produced a similar proportion of SSEA1$^+$CD61$^+$ cells as cultures receiving no tamoxifen (Extended Data Fig. 4g, h). Therefore, OTX2 does not impair the function of an established PGC network, but rather restricts the efficiency with which EpiLCs enter the germline.

PGCLC induction is considered to strictly require cytokine addition[12]. Consistent with this, Oct4ΔPE::GFP was not expressed by aggregates of *Otx2*$^{fl/-}$ cells cultured in the absence of cytokines (Fig. 3a). However, in *Otx2*$^{-/-}$ cells, cytokines were not essential for OctΔPE reporter activation (Fig. 3a), CD61 and SSEA1 surface expression (Fig. 3b, Extended Data Fig. 5a, b) or PGC transcription factor expression (Extended Data Fig. 5c, d). Indeed, mRNA profiling indicated that *Otx2*$^{+/+}$ and *Otx2*$^{-/-}$ EpiLCs were transcriptionally similar and that following differentiation, *Otx2*$^{-/-}$ cells resembled wild-type PGCLCs, irrespective of their exposure to cytokines (Extended Data Fig. 6a). Principal component and ternary analysis confirmed these assessments (Fig. 3c, Extended Data Fig. 6b). These results indicate that without OTX2, germline entry does not require cytokines.

BLIMP1 is essential for wild-type cells to access the germline[3,4]. To determine whether *Otx2*$^{-/-}$ cells retained this dependency, both *Blimp1* alleles were deleted from ESCs of distinct *Otx2* genotypes using CRISPR–Cas9 (Extended Data Figs. 1, 7a–c). PGCLC differentiation confirmed a BLIMP1 requirement for germline entry of OTX2-expressing cells (Fig. 3d, Extended Data Fig. 7d). However, deletion of *Blimp1* from *Otx2*$^{-/-}$ cells did not affect the ability of *Otx2*$^{-/-}$ cells to induce CD61 and SSEA1 (Fig. 3d, Extended Data Fig. 7d). Although deletion of *Blimp1* from wild-type cells increased expression of somatic transcripts during PGCLC differentiation, this did not occur in *Otx2*$^{-/-}$*Blimp1*$^{-/-}$ PGCLCs (Extended Data Fig. 8a). Nor did deletion of *Blimp1* impair the enhanced ability of *Otx2*$^{-/-}$ cells to activate expression of *Prdm14*, *Ap2γ*, *Nanog* or *Oct4* mRNAs (Extended Data Fig. 8b) or DAZL protein (Extended Data Fig. 8c). During differentiation, PGCLCs that express OCT4 have higher H3K27me3 and lower H3K9me2 than OCT4-low cells. These relationships were maintained in *Otx2*$^{-/-}$ and *Otx2*$^{-/-}$*Blimp1*$^{-/-}$ PGCLCs (Extended Data Fig. 8d). Furthermore, without PGCLC cytokines, CD61/SSEA1 expression and Oct4ΔPE reporter activation were unaffected by *Blimp1* deletion

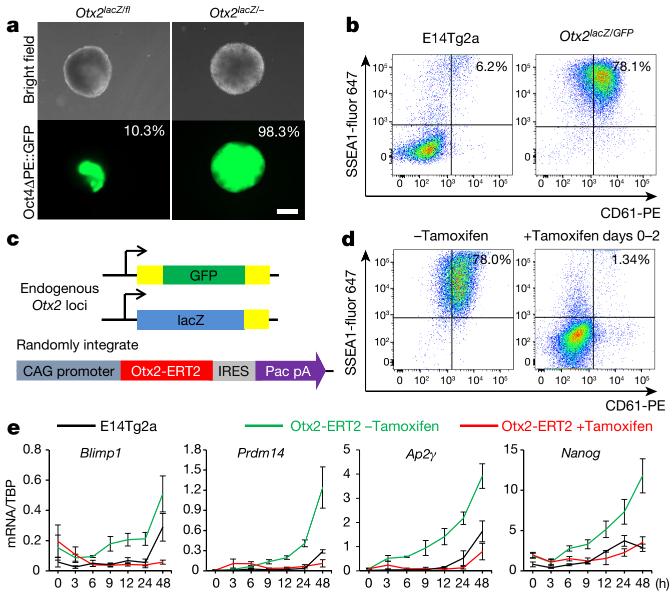


Fig. 2 | *Otx2*$^{-/-}$ EpiLCs have an enhanced propensity to differentiate into PGCLCs. a, Representative morphologies and Oct4ΔPE::GFP expression in aggregates at day 4 of PGCLC differentiation in the presence of cytokines. Percentages indicate GFP-positive cells; $n=9$; scale bar, 200 µm. **b**, E14Tg2a and *Otx2*$^{-/-}$ cells were assessed by flow cytometry for surface expression of SSEA1 and CD61 at days 6 of PGCLC differentiation; $n=12$. **c**, Diagram of the tamoxifen-inducible *Otx2* cell line, carrying an Otx2-ER^{T2} fusion protein transgene and replacements of endogenous *Otx2* alleles by *GFP* or *LacZ*. **d**, *Otx2*$^{lacZ/GFP}$::Otx2ER^{T2} cells were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of PGCLC differentiation, either without tamoxifen or with tamoxifen from days 0–2; $n=4$ biologically independent replicates. **e**, RT–PCR of PGC transcription factors. Expression levels are normalized to TBP; values are means ± s.d.; $n=3$ biologically independent replicates.

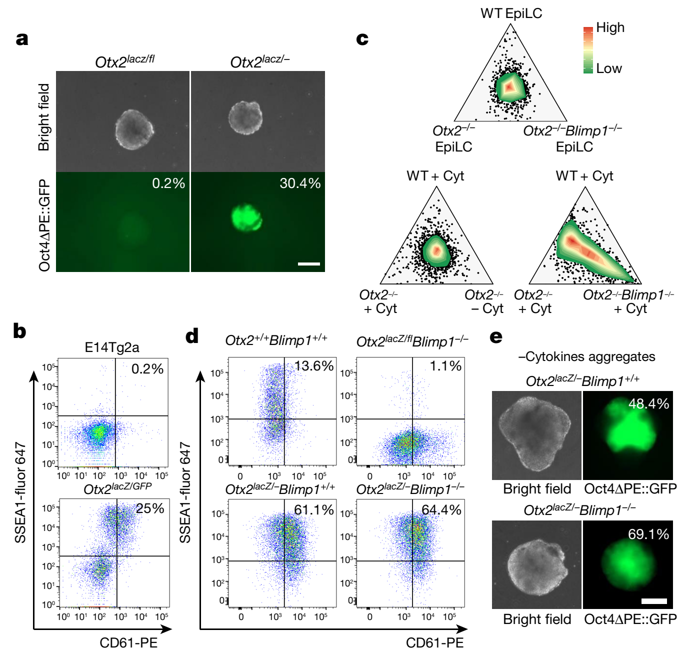


Fig. 3 | *Otx2*-null cells can access the germline independently of cytokines and BLIMP1. a, Representative morphologies and Oct4ΔPE::GFP expression in aggregates at day 4 of PGCLC differentiation without cytokines. Percentages indicate GFP-positive cells; ($n=7$); scale bar, 200 µm. **b**, E14Tg2a and *Otx2*$^{-/-}$ cells were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of differentiation in the absence of PGCLC cytokines; ($n=9$). **c**, Ternary plot analysis (microarray data from three biologically independent replicates under seven different conditions) comparing transcriptomes of EpiLCs (top) or day 6 PGCLCs (bottom). Circles represent probes, with colour indicating the probe density. Differentiations performed in the presence or absence of cytokines are indicated (+/− cyt). WT, E14Tg2a; O$^{-/-}$, *Otx2*$^{lacZ/GFP}$; O$^{-/-}$ B$^{-/-}$, *Otx2*$^{lacZ/GFP}$ *Blimp1*$^{-/-}$. **d**, Cells of the indicated genotypes were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of PGCLC differentiation; $n=2$ for 1 clone of each genotype. Two further clones of each genotype are shown in Extended Data Fig. 7. **e**, Cytokine-free PGCLC differentiation. Representative morphologies and Oct4ΔPE::GFP expression of aggregates at day 6. Percentages indicate GFP-positive cells; $n=2$ for one clone of each genotype; scale bar, 200 µm.

in multiple *Otx2*$^{-/-}$ cell lines (Fig. 3e, Extended Data Fig. 7e, f). Nevertheless, at day 6 of PGCLC differentiation, *Otx2*$^{-/-}$*Blimp1*$^{-/-}$ cells were unable to adopt the mature PGCLC transcriptome observed in *Otx2*$^{+/+}$ or *Otx2*$^{-/-}$*Blimp1*$^{+/+}$ cells (Fig. 3c, Extended Data Fig. 6). These results indicate that without OTX2, BLIMP1 is not required for phenotypic aspects of germline entry, but is required for a fully mature PGC phenotype.

Although EpiLCs respond to cytokine induction by PGCLC differentiation, this is not maintained upon continued passaging in medium containing Activin/FGF[12], suggesting that in these conditions, cells lose germline competence. To assess whether OTX2 affects the period during which pluripotent cells remain competent for germline entry, EpiLCs were passaged in EpiLC medium for a further two days (Fig. 4a). At this point the Oct4ΔPE::GFP reporter is inactive in both *Otx2*$^{fl/-}$ and *Otx2*$^{-/-}$ cells (Extended Data Fig. 9a). Cells were then transferred to PGCLC differentiation medium for 6 days. Notably, Oct4ΔPE::GFP was reactivated robustly in *Otx2*$^{-/-}$ but not *Otx2*$^{fl/-}$ cells (Fig. 4b). Moreover, whereas all cell lines expressed *Blimp1* and *Ap2γ* mRNAs, CD61/SSEA1 surface expression and *Prdm14* and *Nanog* mRNA expression were observed only in *Otx2*$^{-/-}$ and not in *Otx2*$^{+/+}$ or *Otx2*$^{fl/-}$ cells (Extended Data Fig. 9b–d). This indicates that in the absence of OTX2 the period of competence to enter the germline is extended.

To determine whether *Otx2*$^{-/-}$ cells exhibit an increased propensity to enter the germline in vivo, *Otx2*$^{+/+}$ or *Otx2*$^{-/-}$ ESCs constitutively expressing GFP were compared in chimaeras following morula aggregation (Fig. 4c). *Otx2*$^{+/+}$ and *Otx2*$^{-/-}$ cells had a comparable capacity to produce chimaeras (Extended Data Fig. 9e, g). However, an enhanced proportion of *Otx2*$^{-/-}$ cells expressed BLIMP1 or SOX2 (Fig. 4d, Extended Data Fig. 9f), indicating that enhanced germline entry is a cell autonomous property of *Otx2*$^{-/-}$ cells.

To determine whether *Otx2*$^{-/-}$ embryos also showed enhanced PGC numbers, *Otx2*$^{+/-}$ mice[15] were inter-crossed and *Otx2*$^{-/-}$ embryos analysed at E7.5. These embryos show strong developmental defects[15,16] but also showed increased PGC numbers (Fig. 4e, Extended Data Fig. 10a, b), confirming that in vivo, OTX2 acts as a negative regulator of the PGC programme.

Previous studies have shown that, in mice, competence to enter the germline exists transiently in embryos from E5.5–E6.25[1]. In wild-type cells, germline entry requires BMP4 signalling from the extraembryonic ectoderm[2] and is critically dependent on the downstream action of BLIMP1[3]. Our work identifies OTX2 as an intermediary fulcrum in these processes (Fig. 4f). During the period of PGC competence, BMP4 represses *Otx2* expression, partly by endogenous Wnt activation (Fig. 4g). This reduction in OTX2 is necessary for expression of the PGC transcription factors BLIMP1, PRDM14, AP2γ and NANOG as enforcing OTX2 activity prevents their expression. Moreover, the rate of Otx2 decline appears important as without cytokines, germline entry is diminished. We propose that without cytokines, the window for germline entry closes before OTX2 is reduced below a threshold necessary for PGC transcription factor expression. Furthermore, in the absence of OTX2, PGC transcription factor expression does not require BMP4, indicating that BMP4 functions by repressing *Otx2*. *Otx2*$^{-/-}$ cells also exhibit an extended competence period, suggesting that OTX2 starts a process that defines the extent of the competence

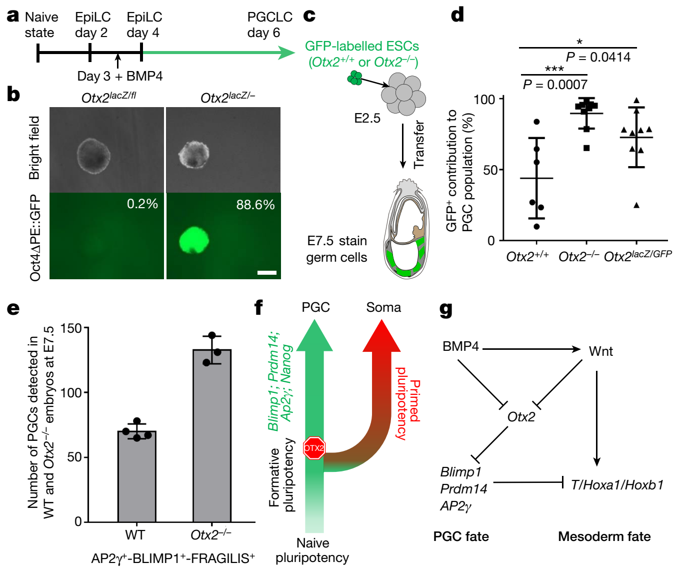


Fig. 4 | *Otx2*$^{-/-}$ ESCs contribute to the germline at an enhanced rate in vivo. a, Scheme for PGCLC differentiation, initiated from day 4 EpiLCs obtained after one passage from EpiLCs. **b**, Representative morphologies and Oct4ΔPE::GFP expression from aggregates at day 6 of PGCLC differentiation from EpiLCs day 4; $n=3$ for 1 clone of each genotype; scale bar, 200 μm. **c**, Scheme for generating chimaeras of GFP-labelled *Otx2*$^{+/+}$ or *Otx2*$^{-/-}$ ESCs with wild-type host embryos. **d**, Comparison of the percentage contribution of GFP-labelled wild-type ($n=6$) or *Otx2*-null ESCs (genotypes indicated, $n=9$ for each) to the PGC population in E7.5 chimaeric embryos. Each dot represents the percentage from one chimaera, centre lines and error bars represent means ± s.d. *P* value (two-sided unpaired *t*-test, 0.95 confidence intervals) is indicated. GFP-positive cells were counted within the PGC population marked with BLIMP1 or SOX2 in each embryo. **e**, Comparison of PGCs number in wild-type ($n=4$) and *Otx2*$^{-/-}$ ($n=3$) E7.5 embryos. PGCs were identified with BLIMP1, AP2γ and FRAGILIS; values are means ± s.d. **f**, Model indicating the point of operation of OTX2 during germline and soma segregation. **g**, A scheme illustrating the regulatory relationships upstream and downstream of *Otx2* during germline segregation.

period. Also, as *Otx2*$^{-/-}$ cells can initiate PGC differentiation without BLIMP1, this suggests a reciprocal relationship between BLIMP1 and OTX2, in which BLIMP1 represses[21,29,30] and OTX2 activates somatic gene expression[17]. Supporting this, during PGCLC differentiation, *Otx2*$^{-/-}$ cells do not activate mesendoderm genes[17] (Extended Data Fig. 4c) that are otherwise repressed by BLIMP1[30]. This may explain why some aspects of the PGCLC differentiation phenotype can be divorced from a BLIMP1 requirement in *Otx2*$^{-/-}$ cells. This places OTX2 at a developmental crossroads where it acts to control excessive access to the germline (Fig. 4f).

Finally, our findings are noteworthy in light of the hypothesis that the neural lineage is the default developmental pathway for vertebrate cells[31]. Interestingly, neural induction requires inhibition of BMP signalling[32]. BMP is a known facilitator of germline entry[2] and is identified here as a key *Otx2* repressor. The default neural induction hypothesis is based principally on studies in chicks and frogs, species in which PGCs are formed by germplasm segregation. Yet, induced germline segregation is considered the ancestral mechanism that predates the recurrent evolution of germplasm[7,8]. The highly efficient entry of pluripotent cells into the germline in the absence of OTX2 reported here suggests that the germline may be the ancient default option that must be overcome in order to elaborate the ancillary structures of the soma.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0581-5.

Received: 7 February 2018; Accepted: 24 August 2018;
Published online 3 October 2018.

1. Ohinata, Y. et al. A signaling principle for the specification of the germ cell lineage in mice. *Cell* **137**, 571–584 (2009).
2. Lawson, K. A. et al. Bmp4 is required for the generation of primordial germ cells in the mouse embryo. *Genes Dev.* **13**, 424–436 (1999).
3. Ohinata, Y. et al. Blimp1 is a critical determinant of the germ cell lineage in mice. *Nature* **436**, 207–213 (2005).
4. Vincent, S. D. et al. The zinc finger transcriptional repressor Blimp1/Prdm1 is dispensable for early axis formation but is required for specification of primordial germ cells in the mouse. *Development* **132**, 1315–1325 (2005).
5. Yamaji, M. et al. Critical function of Prdm14 for the establishment of the germ cell lineage in mice. *Nat. Genet.* **40**, 1016–1022 (2008).
6. Weber, S. et al. Critical function of AP-2 gamma/TCFAP2C in mouse embryonic germ cell maintenance. *Biol. Reprod.* **82**, 214–223 (2010).
7. Johnson, A. D. & Alberio, R. Primordial germ cells: the first cell lineage or the last cells standing? *Development* **142**, 2730–2739 (2015).
8. Extavour, C. G. & Akam, M. Mechanisms of germ cell specification across the metazoans: epigenesis and preformation. *Development* **130**, 5869–5884 (2003).
9. McLaren, A. Primordial germ cells in the mouse. *Dev. Biol.* **262**, 1–15 (2003).
10. Surani, M. A., Hayashi, K. & Hajkova, P. Genetic and epigenetic regulators of pluripotency. *Cell* **128**, 747–762 (2007).
11. Saitou, M. & Yamaji, M. Primordial germ cells in mice. *Cold Spring Harb. Perspect. Biol.* **4**, a008375 (2012).
12. Hayashi, K., Ohta, H., Kurimoto, K., Aramaki, S. & Saitou, M. Reconstitution of the mouse germ cell specification pathway in culture by pluripotent stem cells. *Cell* **146**, 519–532 (2011).
13. Buecker, C. et al. Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14**, 838–853 (2014).
14. Yang, S. H. et al. Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency. *Cell Reports* **7**, 1968–1981 (2014).
15. Acampora, D. et al. Forebrain and midbrain regions are deleted in Otx2$^{-/-}$ mutants due to a defective anterior neuroectoderm specification during gastrulation. *Development* **121**, 3279–3290 (1995).
16. Ang, S. L. et al. A targeted mouse Otx2 mutation leads to severe defects in gastrulation and formation of axial mesoderm and to deletion of rostral brain. *Development* **122**, 243–252 (1996).
17. Acampora, D., Di Giovannantonio, L. G. & Simeone, A. Otx2 is an intrinsic determinant of the embryonic stem cell state and is required for transition to a stable epiblast stem cell condition. *Development* **140**, 43–55 (2013).
18. Acampora, D. et al. Functional antagonism between OTX2 and NANOG specifies a spectrum of heterogeneous identities in embryonic stem cells. *Stem Cell Reports* **9**, 1642–1659 (2017).
19. Nakaki, F. et al. Induction of mouse germ-cell fate by transcription factors *in vitro*. *Nature* **501**, 222–226 (2013).
20. Murakami, K. et al. NANOG alone induces germ cells in primed epiblast *in vitro* by activation of enhancers. *Nature* **529**, 403–407 (2016).
21. Magnúsdóttir, E. et al. A tripartite transcription factor network regulates primordial germ cell specification in mice. *Nat. Cell Biol.* **15**, 905–915 (2013).
22. Smith, A. Formative pluripotency: the executive phase in a developmental continuum. *Development* **144**, 365–373 (2017).
23. Günesdogan, U. & Surani, M. A. Developmental competence for primordial germ cell fate. *Curr. Top. Dev. Biol.* **117**, 471–496 (2016).
24. Saitou, M., Barton, S. C. & Surani, M. A. A molecular programme for the specification of germ cell fate in mice. *Nature* **418**, 293–300 (2002).
25. Yoshimizu, T. et al. Germline-specific expression of the Oct-4/green fluorescent protein (GFP) transgene in mice. *Dev. Growth Differ.* **41**, 675–684 (1999).
26. Zhang, M. et al. Esrrb complementation rescues development of Nanog-null germ cells. *Cell Reports* **22**, 332–339 (2018).
27. Aramaki, S. et al. A mesodermal factor, T, specifies mouse germ cell fate by directly activating germline determinants. *Dev. Cell* **27**, 516–529 (2013).
28. Ben-Haim, N. et al. The nodal precursor acting via activin receptors induces mesoderm by maintaining a source of its convertases and BMP4. *Dev. Cell* **11**, 313–323 (2006).
29. John, S. A. & Garrett-Sinha, L. A. Blimp1: a conserved transcriptional repressor critical for differentiation of many tissues. *Exp. Cell Res.* **315**, 1077–1084 (2009).
30. Kurimoto, K. et al. Complex genome-wide transcription dynamics orchestrated by Blimp1 for the specification of the germ cell lineage in mice. *Genes Dev.* **22**, 1617–1635 (2008).
31. Hemmati-Brivanlou, A. & Melton, D. Vertebrate embryonic cells will become nerve cells unless told otherwise. *Cell* **88**, 13–17 (1997).
32. Levine, A. J. & Brivanlou, A. H. Proposal of a model of mammalian neural induction. *Dev. Biol.* **308**, 247–256 (2007).

Acknowledgements We thank V. Wilson and D. O'Carroll for comments on the manuscript, V. Wilson for help with embryo staging, N. Mullin for pre-mRNA analyses, P. Moreira for help with embryo transfer, the CRM animal house staff for husbandry, F. Rossi and C. Cryer for FACS and B. Vernay for confocal assistance. This research was funded by the Medical and the Biotechnological and Biological Sciences Research Councils of the UK (I.C.), by a PRIN project from MIUR (A.S.) and by the Qilu Young Scholars Program of Shandong University (D.Y.).

Reviewer information *Nature* thanks K. Hayashi, A. Johnson and D. Laird for their contribution to the peer review of this work.

Author contributions J.Z., M.Z. and I.C. conceived the project and designed experiments. A.S. and D.A. analysed *Otx2*-null embryos and provided reagents. D.Y. performed bioinformatics analysis. J.Z. and M.Z. performed experiments, with help from M.V. J.Z., M.Z. and I.C. analysed the data. I.C. wrote the paper with input from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0581-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0581-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to I.C. or M.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment. All mouse studies were performed in accordance with UK Home Office regulations under project licence PPL 60/4435 and carried out in a Home Office-designated facility.

Cell culture and differentiation. ESCs were routinely cultured in GMEMβ supplemented with 100 U ml^{-1} LIF and 10% FCS at $3–10 \times 10^5$ cells per cm^2 density on tissue culture flasks coated with 0.1% gelatin[33]. Cells were routinely tested for mycoplasma contamination. EpiLC differentiation and PGCLC induction were carried out according to protocol previously described[12,26,34].

For cytokine-free differentiation, EpiLCs were dissociated and resuspended at 8×10^4 cells ml^{-1} in GK15 medium (GMEMβ supplemented with 15% KOSR). 25 μl drops containing 2,000 cells were plated on the lids of tissue culture dishes and incubated over a reservoir of PBS for 2 days. Hanging drops were then collected and transferred to untreated culture dishes supplemented with GK15 medium and rotated at 72 r.p.m. Fresh medium was replenished every other day. CHIR (3 μM) and XAV939 (2 μM) were used to induce and repress Wnt signalling.

Generation of knockout cell lines. For a summary of cell lines used in this report and a history of their derivation, see Extended Data Figure 1.

Otx2-knockout (*Otx2*$^{lacZ/GFP}$) and *Otx2* conditional knockout (*Otx2*$^{lacZ/fl}$; *Rosa26*$^{CreER/+}$) cell lines and targeting strategies have been described previously[17]. To derive *Otx2*-knockout (*Otx2*$^{lacZ/-}$) lines from *Otx2*$^{lacZ/fl}$, *Otx2*$^{lacZ/fl}$ ESCs were treated with 1 μM tamoxifen for 24 h, plated at clonal density and after 6 days, individual colonies picked, expanded and genotypes verified with RT-PCR, western blot and immunostaining.

For knockout cell lines derived using CRISPR/Cas9 technology, ESCs were transfected with vectors expressing gRNAs and eSpCas9-T2A-mCherry or eSpCas9-T2A-eGFP using Lipofectamine 3000 (Invitrogen, L3000015) and cultured for 24 h. After FACS sorting, single-cell clones were expanded and genotyped. Clones with predicted genotypes were further verified with qPCR or western blot and immunostaining. See Supplementary Table 1 for detailed oligonucleotide sequences of gRNAs, genotyping PCR.

Immunohistochemistry. For whole-mount staining of embryos and PGCLC aggregates, embryos and PGCLC aggregates were fixed with 4% paraformaldehyde (PFA) in PBS (room temperature, 30 min). For Fragilis staining, embryos were washed three times in PBS/0.1%BSA and blocked overnight at 4 °C in 3% donkey serum (Sigma)/1% BSA/PBS solution. For other antibodies, samples were washed three times with PBS/0.1% TritonX100 (PBST), permeabilized in 0.5% Triton X100/PBS solution for 15 min and incubated in 1M glycine in PBST for 20 min. After three washes, samples were blocked overnight at 4 °C using 3% donkey serum, 1% BSA (Sigma) in PBST (blocking buffer). Samples were then incubated for 72 h with diluted primary antibodies, rinsed four times for 20 min each in PBST and incubated for 4 h at room temperature with diluted secondary antibody. Samples were then rinsed four times for 20 min each in PBST. DAPI (Molecular Probes, D1306) was used for nuclear staining (same for other staining). For imaging, embryos were treated with 10%, 25%, 50%, 97% thiodiethanol (Sigma, 166782) for 5 min in each gradient and were imaged using the TCS SP8 inverted confocal microscope (Leica).

For staining of frozen sections, aggregations were fixed in 4% PFA (room temperature, 20 min). After washing in PBS, samples were embedded in Tissue-Freezing medium (Thermo Fisher scientific, 6502Y) and sectioned at a thickness of 5 μm. For antigen retrieval, sections were microwaved with highest power in 10 mM sodium citrate (PH, 6.0) for 2 min 30 s, twice. After antigen retrieval, samples were incubated in blocking buffer (room temperature, 1 h), then incubated with primary antibodies (4 °C overnight), after being rinsed three times in PBST for 10 min, incubated for 2 h at room temperature with diluted secondary antibody. Sections were then covered by cover slip in Fluoromount (SouthernBiotech, 0100-01) and imaged using the TCS SP8 inverted confocal microscope (Leica).

For immunostaining of cells grown in monolayer, cells were washed once with PBS, fixed in 4% PFA (room temperature, 10 min), permeabilized in 0.3% Triton X100/PBS (room temperature, 20 min), incubated in blocking buffer (room temperature, 1 h), before addition of primary antibodies and incubation at 4 °C overnight. Cells were washed in PBST (four times, 5 min) before incubation with diluted secondary antibodies (room temperature, 1 h). After washing four times (PBST), cells were imaged using the TCS SP8 inverted confocal microscope (Leica).

For cytospin staining, cells or aggregates were dissociated into single cells. 1×10^5 cells in 100 μl 1% BSA/PBS buffer were added into sample holders, centrifuged in the cytospin machine (5 min, 1200 r.p.m.). Microscope slides were then taken from the holders and cells circled with hydrophobic marker pen. Staining was then finished following the same protocol as for monolayer cells.

For PGC cell counting in wild-type and *Otx2*-null embryos, embryos were isolated at E7.5, washed in PBS, fixed overnight in 4% PFA/PBS, dehydrated and paraffin embedded as previously reported[18]. Embryos were sectioned in coronal-frontal or sagittal sections and processed for immunohistochemistry (IHC) with antibodies against BLIMP1, AP2γ (Santa Cruz Biotechnology, SC-53162) and Fragilis (R&D Systems, AF3377). All sections were analysed and those including PGCs captured for cell-counting analysis.

If not specified, primary and secondary antibodies used are listed in Supplementary Table 2.

Immunofluorescence quantification. Cytospin slides were stained at the same time and imaged using a Zeiss Observer microscope (Zeiss), Plan-Apo 20× NA, 0.8 objective (Zeiss), a Hamamatsu ORCA-Flash4.0 V3 camera (Hamamatsu), a Colibri 7 (Zeiss) light source and the following filter cubes (name, excitation LED, beam splitter, band pass emission filter; 49 DAPI, 395, 480/40; 38 HE GFP, 495, 525/50; 43 HE dsRed, 570, 605/70; Cy5, 660, 700/775). Images were analysed using CellProfiler software (version 2.2.0)[35]. DAPI staining was first used to segment individual nuclei based on the diameter and intensity of the objects (25–90 pixel units and intensity threshold >0.05 respectively), then the intensity of OTX2 and AP2γ of the segmented nuclei were measured and the mean intensity of each channel reported. Over 8,000 segmented nuclei of each samples were analysed. The data were analysed in R and plots were generated using the ggplot2 package. The *Otx2*-knockout EpiLCs are negative for both OTX2 and AP2γ staining and therefore were used to set up the threshold for gating OTX2 and AP2γ for all samples (negative gate > 99.5%).

Flow cytometry. FACS analysis was performed as described[26]. Cells grown in monolayer or embryoid bodies in suspension were dissociated into single cells with trypsin and neutralized in PBS/10%FCS. A maximum of 5×10^5 cells were collected by centrifugation and the pellet resuspended in 100 μl PBS/10%FCS supplemented with Alexa Fluor 647 anti-mouse/human CD15 (SSEA-1) (Biolegend, 125608) and PE anti-mouse/rat CD61 (Biolegend, 104307) diluted 1/200 and 1/500, respectively, and incubated (30 min, 4 °C). Cells were washed twice in 1 ml PBS/10% FCS before analysis on a BD Fortessa 5 laser LSRII. Gate strategy is shown in Extended Data Fig. 3a.

SDS–PAGE electrophoresis and immunoblotting. Immunoblot analysis was performed as described[36]. Briefly, protein samples from cell lysates, along with protein ladder (Novex, cat. LC5925) were loaded on 10% Bis-Tris Gels (Novex, cat. BG00102BOX) and electrophoresis performed at 200 V for 60–80 min. Proteins were then transferred onto Nitrocellulose membrane (Capitol Scientific, cat. 10401396), blocked in 10% milk (w/v) (room temperature, 1 h), followed by incubation with primary antibodies (4 °C, overnight). After three washes in PBST, membranes were incubated with IRDye conjugated secondary antibody, followed by three washes in PBST before being visualized in LI-COR Odyssey Imaging Systems. Primary and secondary antibodies are listed in Supplementary Table 2.

RNA analysis. Total RNA was extracted using either Trizol (Invitrogen,15596026) or RNeasy micro (Qiagen, 74034) or mini kit (Qiagen, 74104) following the manufacturer's instructions, followed by DNase treatment (Qiagen, 157047207). Reverse transcription reactions were performed using Superscript First Strand Synthesis kit (Invitrogen, 1964419) and quantitative PCR performed using SYBR Green Kits (Takyon, UF-NSMT-B0701) on Roche LightCycler 480 with cDNA equivalent of 25 ng total RNA per reaction. Values for each gene were normalized to expression of TATA-box Binding Protein (TBP) according to $2^{-\Delta Ct}$ formula[37]. Oligonucleotide sequences are shown in Supplementary Table 3.

Transcriptomic profiling and data analysis. Total RNAs were extracted using RNeasy Plus Micro Kit (Qiagen, cat.74034), following manufacturer's instructions. RNAs were labelled using Illumina TotalPrep RNA Amplification Kit (Illumina, cat. AMIL1791) following manufacturer's instructions. Briefly, 100 to around 300 ng total RNA were reverse transcribed using oligo(dT) primers, followed by second-strand cDNA synthesis. The dsDNAs were then in vitro transcribed into biotin labelled complementary RNA (cRNA) at 37 °C for 12 h. The cRNAs were purified and quality analysed using Agilent Bioanalyzer 2100. The hybridizations were performed by the Eurofins Genomics AROS in Denmark. The samples were hybridized on Illumina MouseWG-6 v2 BeadChip. A signal intensity (expression) value and a detection *P* value (whether the signal is above background) for each probe on the array was obtained using Illumina GenomeStudio software with standard settings. A total of 13,683 probes which had detection P value < 0.01 in at least three experiments were considered for further analysis. A variance-based filtering of gene expressions was performed using the genefilter package with variance cutoff of 0.75, leaving 3,421 probes for further statistical analysis. Unsupervised hierarchy clustering, generation of heatmap and PCA and ternary plot were carried out using R (https://www.R-project.org) using the following packages: cluster, pheatmap, FactoMineR, ggtern and RColorBrewer. The microarray data has been deposited in the GEO database under accession number GSE116640.

Embryo aggregation. E2.5 embryos were collected from superovulated F1 (CBA male cross with C57BL/6J female) females with M2 buffer (Sigma, M1767). After being cultured in KSOM (Millipore, MR-020P-5F) for 15 min, zonae pellucidae were removed by Tyrode's solution (Sigma, T1788) and embryos cultured in KSOM in aggregation plates prepared by aggregation needle (BLS, DN-10). Wild-type

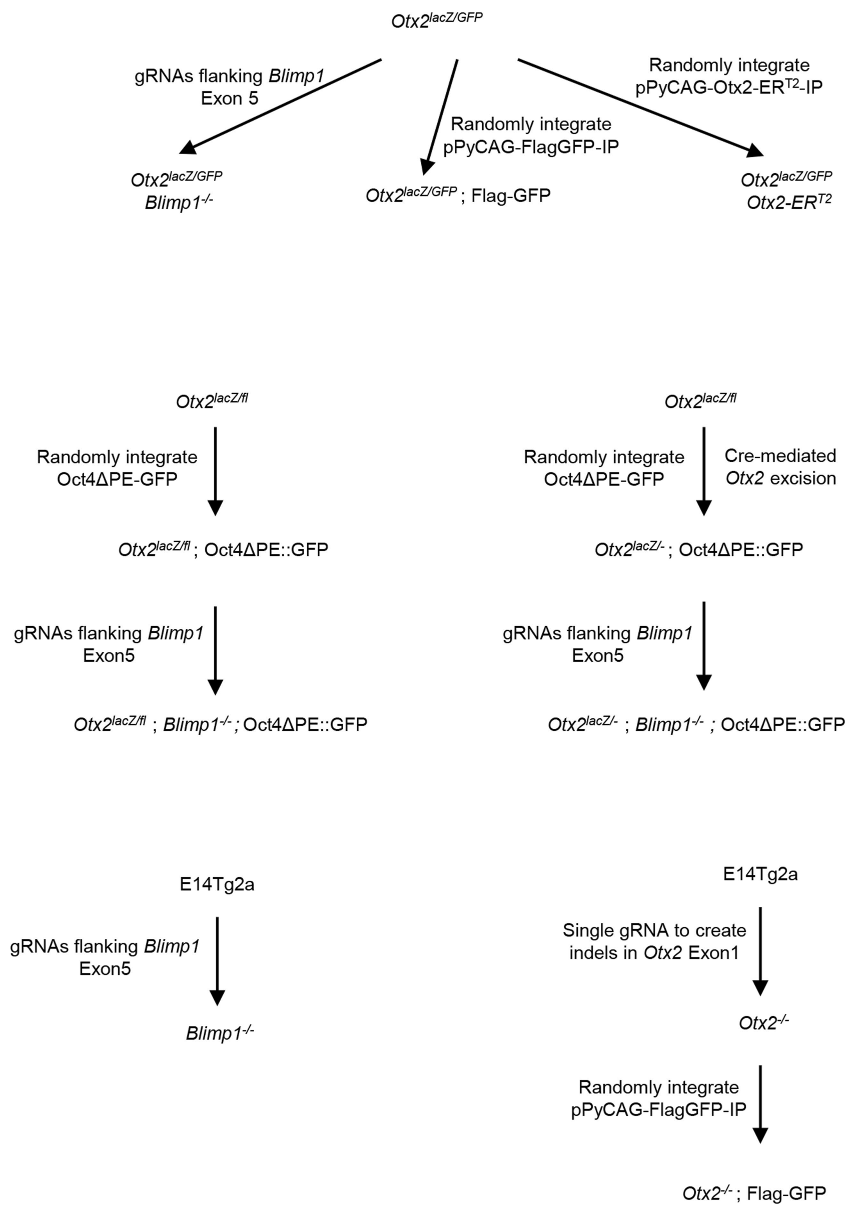
E14Tg2a ESCs and two independent *Otx2*-knockout cell lines (labelled with Flag–GFP) were used for aggregation. Cells were trypsinized (37 °C, 1 min), washed and re-suspended with GMEM/FCS medium. Each embryo was aggregated with 6–8-cell clumps and cultured in the incubator. The next day, good-quality blastocysts were picked and transferred to E0.5 CD1 recipient[38].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

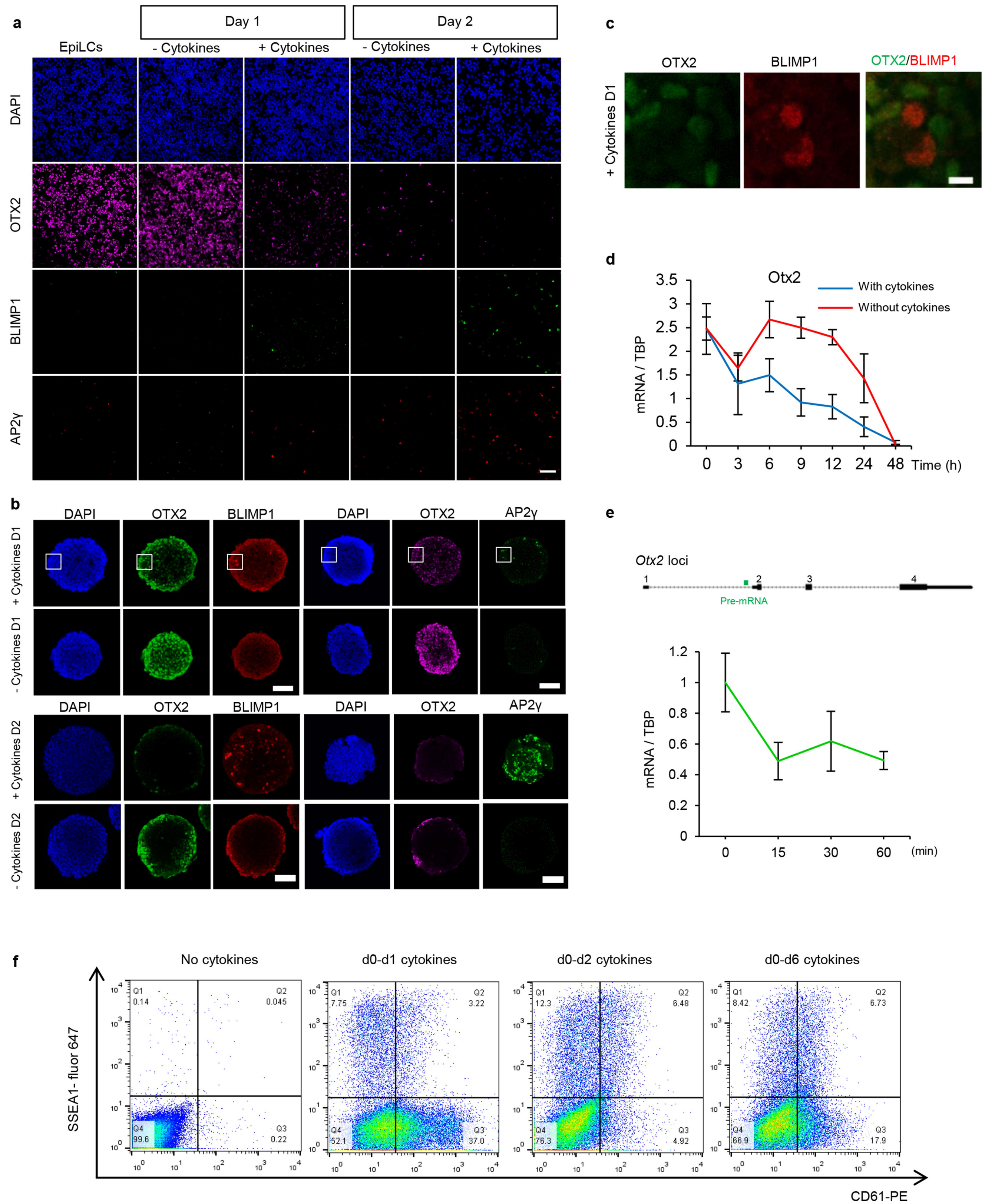
Data availability

All the data sets generated or analysed during the current study are available from the corresponding author on reasonable request.

33. Smith, A. G. Culture and differentiation of embryonic stem cells. *J. Tissue Cult. Methods* **13**, 89–94 (1991).
34. Hayashi, K. & Saitou, M. Generation of eggs from mouse embryonic stem cells and induced pluripotent stem cells. *Nat. Protocols* **8**, 1513–1524 (2013).
35. Lamprecht, M. R., Sabatini, D. M. & Carpenter, A. E. CellProfiler: free, versatile software for automated biological image analysis. *Biotechniques* **42**, 71–75 (2007).
36. Gagliardi, A. et al. A direct physical interaction between Nanog and Sox2 regulates embryonic stem cell self-renewal. *EMBO J*. **32**, 2231–2247 (2013).
37. Schmittgen, T. D. & Livak, K. J. Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protocols* **3**, 1101–1108 (2008).
38. Bronson, R. A. & McLaren, A. Transfer to the mouse oviduct of eggs with and without the zona pellucida. *J. Reprod. Fertil*. **22**, 129–137 (1970).

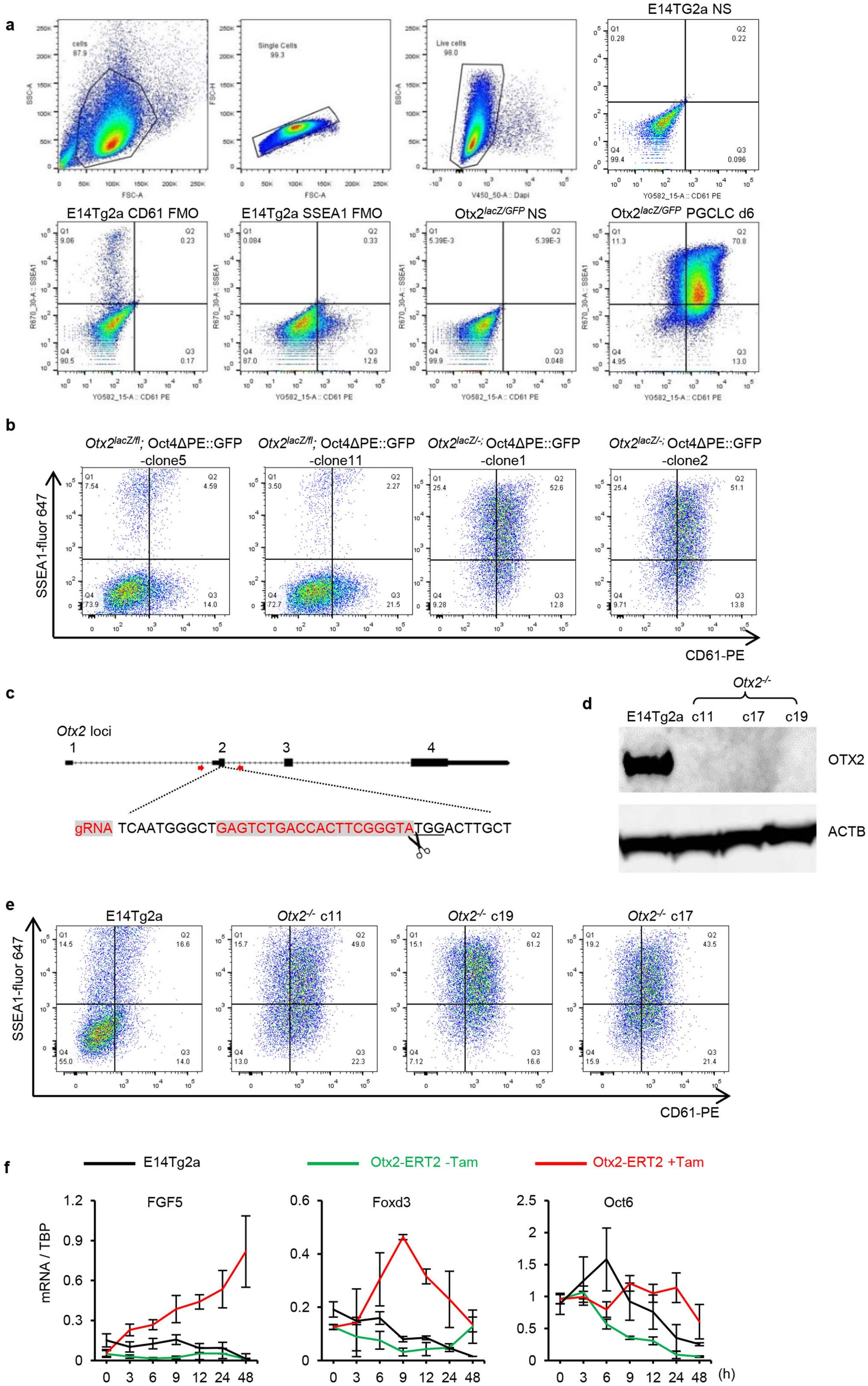


Extended Data Fig. 1 | Summary of cell lines used in this report. *Otx2*$^{lacZ/GFP}$ and *Otx2*$^{lacZ/fl}$ ESCs have been described previously[17]. Summarized below are further modifications to *Otx2* or *Blimp1*, or transgene additions in the above or wild-type backgrounds. Further schematic details illustrating the points of Cas9 modification of *Otx2* or *Blimp1* and genotype verification of derived cell lines are shown in Extended Data Fig. 3 and Extended Data Fig. 7, respectively.



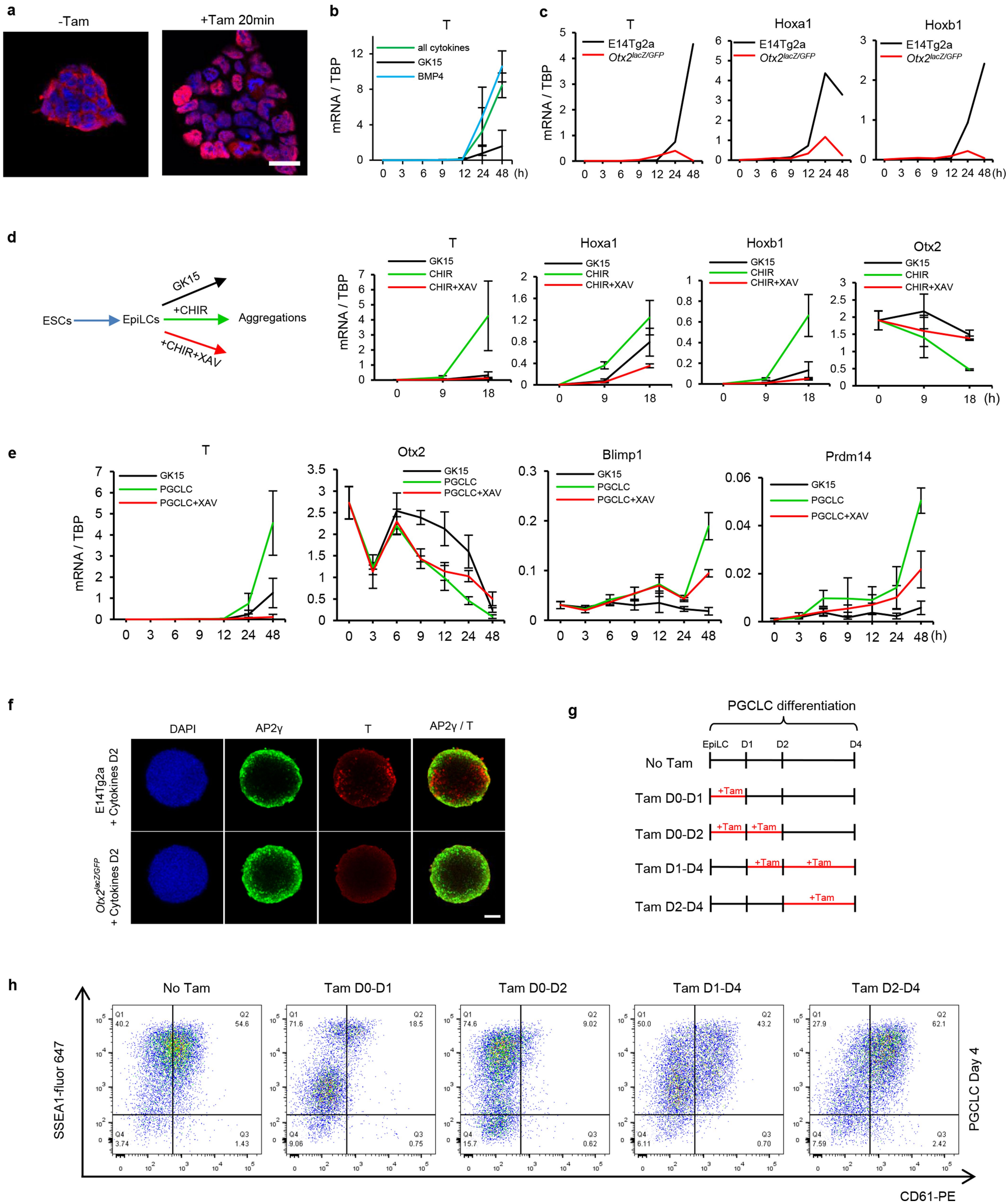
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Competence for germline entry is preceded by downregulation of OTX2 protein. a, Representative cytospin images of OTX2, BLIMP1 and AP2γ staining using E14Tg2a aggregates after 1 day or 2 days of PGCLC differentiation; $n = 2$; scale bar, 100 μm. **b**, Whole-mount immunofluorescence of E14Tg2a aggregates after 1 (D1) or 2 days (D2) of differentiation of EpiLCs in the presence or absence of cytokines. Representative images of OTX2 and BLIMP1 are shown; $n = 3$; scale bar, 50 μm. **c**, Magnified image of the region highlighted in **b**; scale bar, 10 μm. **d**, Quantitative transcript analysis of *Otx2* in E14Tg2a cultures with ($n = 4$) or without cytokines ($n = 7$) at indicated time point. Schematic illustration is shown in Fig. 1b. Expression levels are normalized to TBP; values are means ± s.d. **e**, Top, primers used for *Otx2* pre-mRNA transcript analysis are shown relative to the primary transcript structure. Bottom, quantitative transcript analysis of *Otx2* pre-mRNA at the indicated times (minutes) after changing E14Tg2a EpiLCs into PGCLC medium. Expression levels are normalized to TBP and shown relative to expression at $t = 0$; values are means ± s.d.; $n = 3$ biologically independent replicates. **f**, Assessing the temporal requirement of cytokine treatment for efficient PGCLC induction. Aggregates of E14Tg2a EpiLCs treated with cytokines for 1 (d0–d1), 2 (d0–d2) or 6 days (d0–d6) were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of PGCLCs differentiation; $n = 3$.



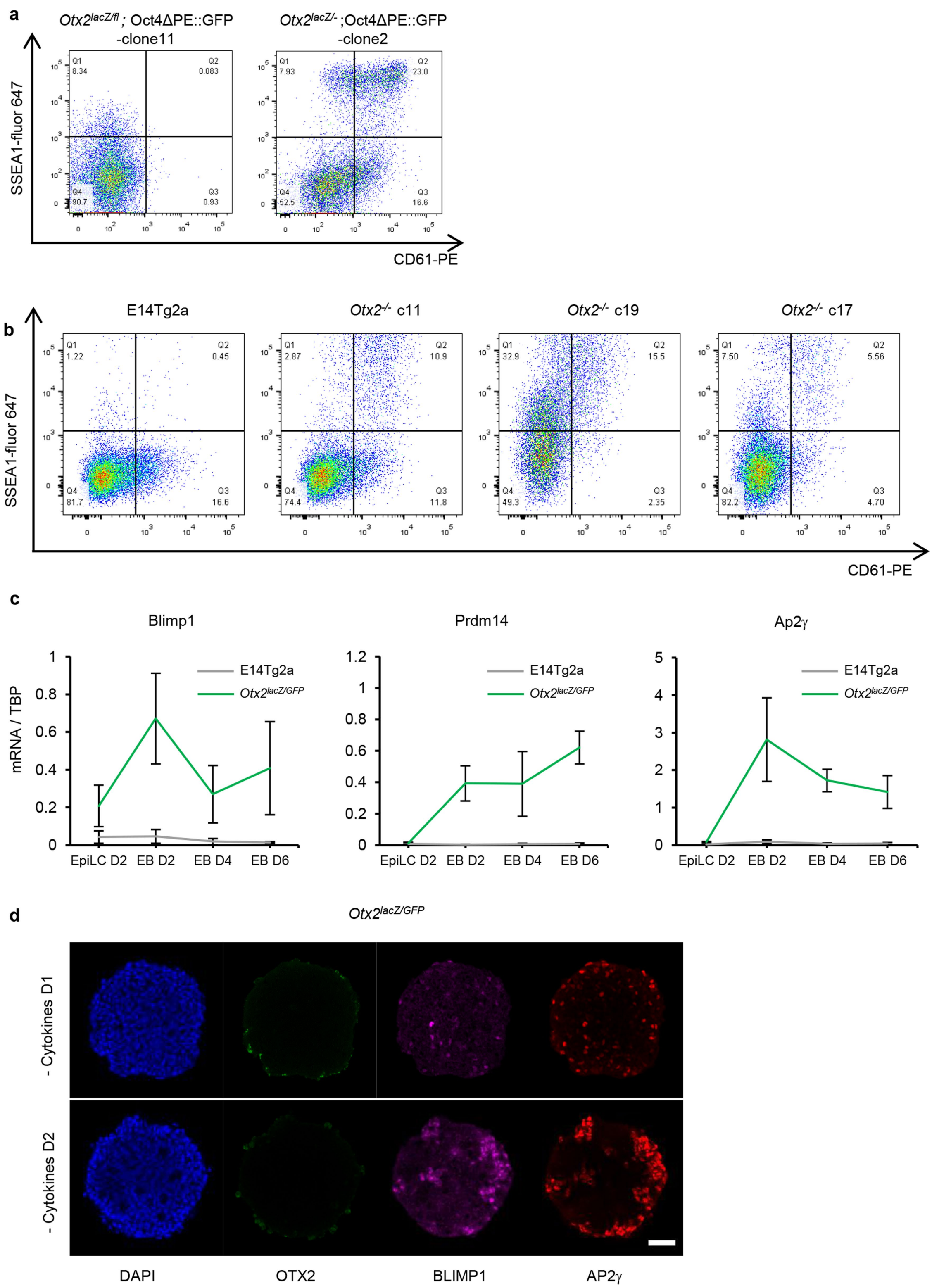
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Independent *Otx2*$^{-/-}$ clones show enhanced PGCLC induction efficiency. a, The gating strategies for analysing PGCLCs by flow cytometry. Cells were first gated based on the FSC (size) and SSC (complexities) scatter plot, followed by selection for singlets based on linear correlations between FSC-area and FSC-height. Live cells were then gated based on exclusion of DAPI to indicate cell membrane integrity. Live cells were then analysed for SSEA1 and CD61. Cells stained for fluorescence minus one (FMO) were used to set gates; stained and non-stained cells are also shown. **b**, *Otx2*$^{lacZ/fl}$ and *Otx2* $^{lacZ/-}$ cells with the Oct4ΔPE::GFP reporter (two independent clones each) were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of PGCLC differentiation. For clone 5 and clone 1, $n = 2$; for clone 11 and clone 2, $n = 9$. **c**, Diagram showing the gRNA sequence (in red) and targeting strategy for generating *Otx2*-knockout cell lines. Red arrows represent genotyping primers used for screening clones. **d**, Immunoblot analysis of OTX2 protein expression in EpiLCs of E14Tg2a and three *Otx2*$^{-/-}$ clones. Experiment preformed once. **e**, E14Tg2a and three independent *Otx2*$^{-/-}$ clones generated by CRISPR/Cas9 were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of PGCLC differentiation. Two biologically independent experiments for clone c11, one for clone c17 and c19. **f**, Q-RT–PCR of epiblast markers during the time-course outlined in Fig. 1b. Expression levels are normalized to TBP; values are means ± s.d.; $n = 3$ biologically independent replicates.

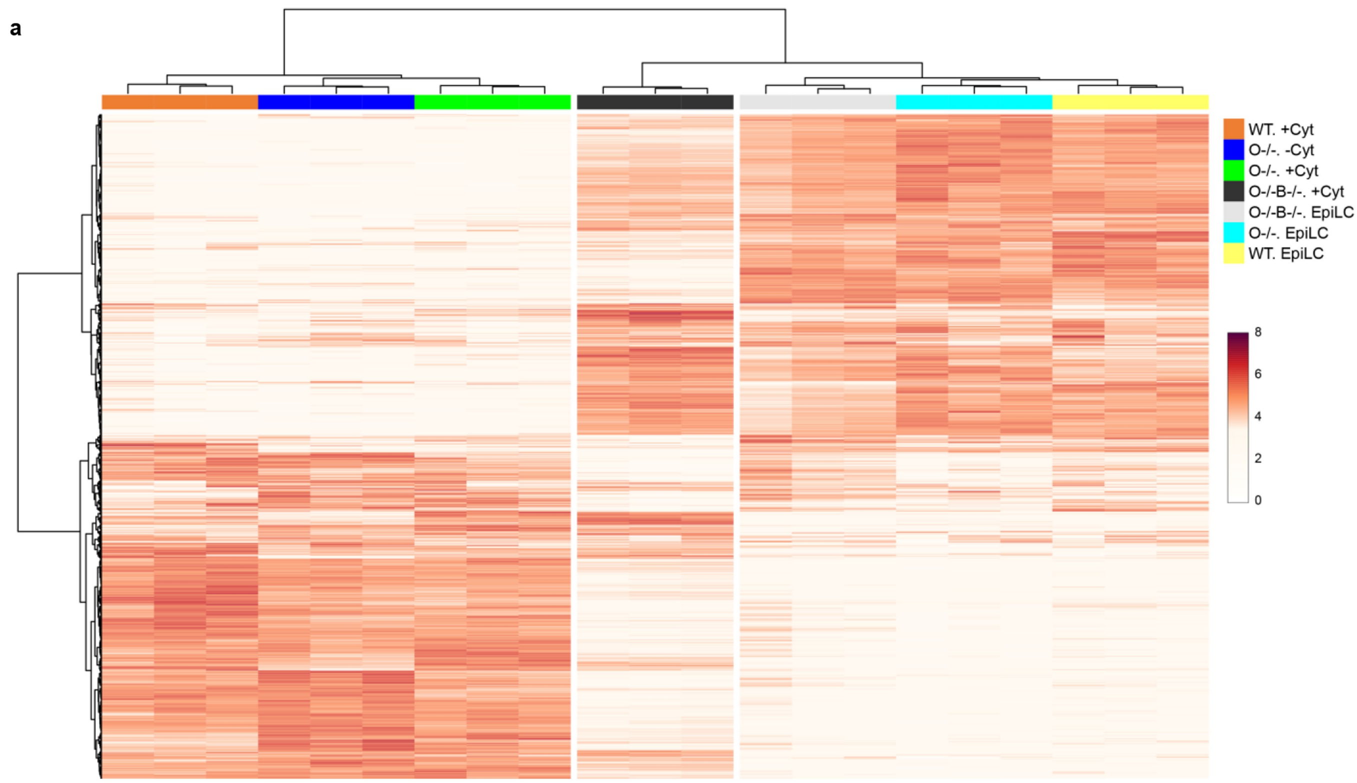


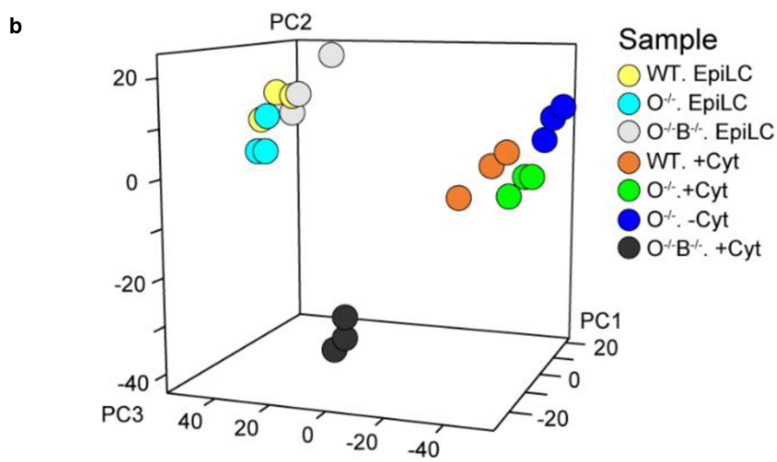
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | OTX2 restricts PGC specification during the first two days of induction. a, OTX2 immunofluorescence of *Otx2*$^{lacZ/GFP}$::Otx2ER^{T2} ESCs before or after treatment with tamoxifen for 20 min; $n=2$ biologically independent experiments; scale bar, 20 μm. **b**, Quantitative transcript analysis of *T* (Brachyury) during the time-course outlined in Fig. 1b in basal GK15 medium supplemented with the indicated cytokines. Expression levels are normalized to TBP; values are means ± s.d.; $n=3$ biologically independent replicates. **c**, Quantitative transcript analysis of *T* (Brachyury), *Hoxa1* and *Hoxb1* during the time-course outlined in Fig. 1b in indicated cell lines. Expression levels are normalized to TBP; values are means from two biologically independent replicates. **d**, Left, scheme illustrating the strategy for induction or repression of Wnt signalling. E14Tg2a EpiLCs were aggregated in the indicated media and transcripts analysed at 0, 9 and 18 h. Right, quantitative transcript analysis of *T* (Brachyury), *Hoxa1*, *Hoxb1* and *Otx2* during the time-courses outlined on the left. Expression levels are normalized to TBP; values are means ± s.d.; $n=3$ biologically independent replicates. **e**, Quantitative transcript analysis of *T* (Brachyury), *Otx2*, *Blimp1* and *Prdm14* during E14Tg2a differentiation in three different media conditions (GK15, without cytokines; PGCLC, GK15 with cytokines; PGCLC +XAV, GK15 with cytokines and with XAV939) at the indicated time point. Expression levels are normalized to TBP; values are means ± s.d.; $n=3$ biologically independent replicates. **f**, Whole-mount immunofluorescence analysis of AP2γ and T (Brachyury) in E14Tg2a and *Otx2*$^{lacZ/GFP}$ day 2 (D2) PGCLC aggregates; $n=2$ biological replicates; scale bar, 50 μm. **g**, Scheme illustrating tamoxifen administration schemes. **h**, *Otx2*$^{lacZ/GFP}$::Otx2ER^{T2} cells were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of PGCLC differentiation following the tamoxifen treatment regime outlined (**g**); $n=2$ biological replicates.

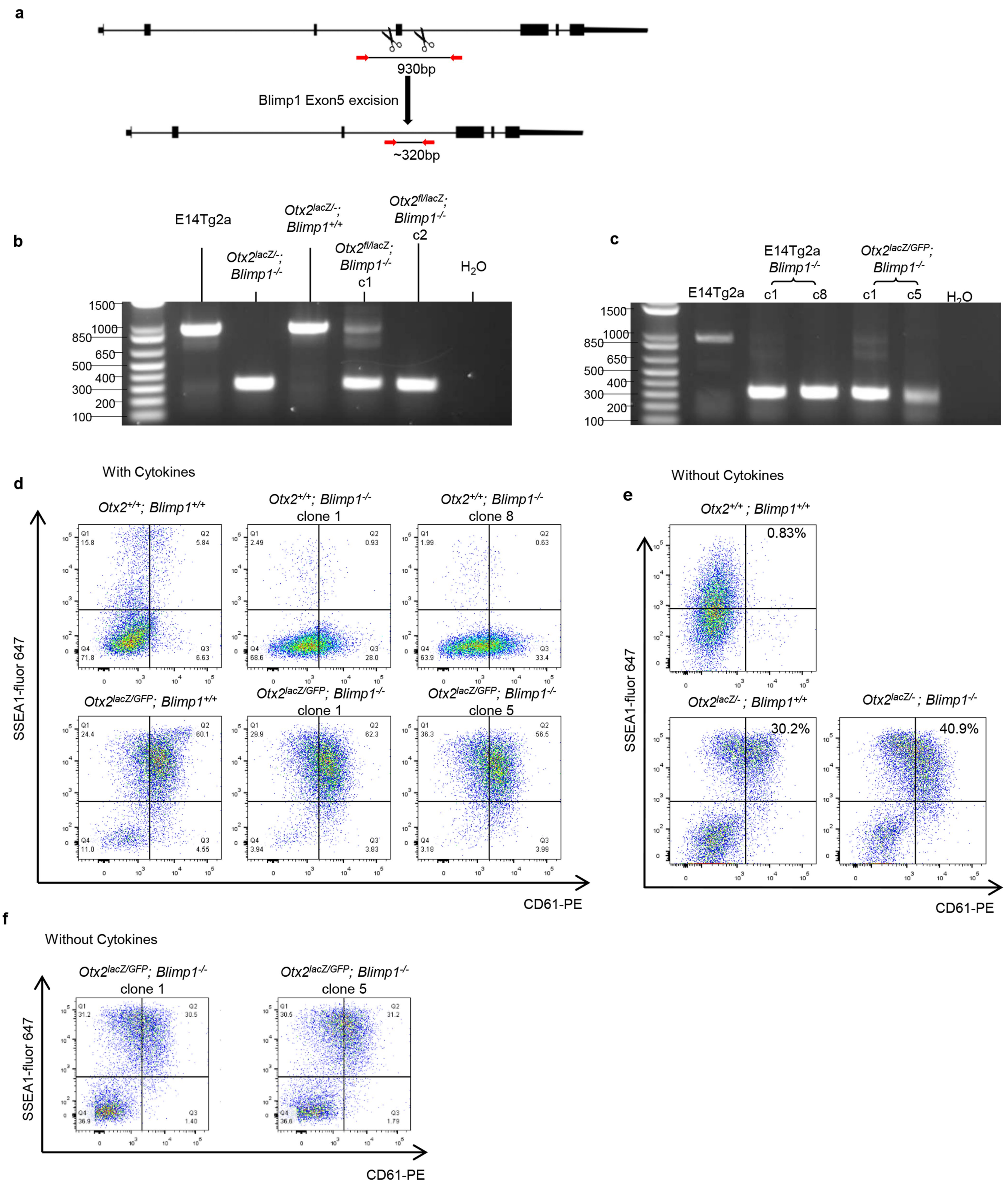


Extended Data Fig. 5 | PGCLC differentiation of *Otx2*-null cells in the absence of cytokines. **a**, *Otx2*$^{lacZ/fl}$ and *Otx2*$^{lacZ/-}$ cells carrying the Oct4ΔPE::GFP reporter (aggregates shown in Fig. 3a) were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of PGCLC differentiation in the absence of cytokines; $n = 7$. **b**, E14Tg2a and three independent *Otx2*$^{-/-}$ clones generated by CRISPR/Cas9 were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of PGCLC differentiation in the absence of cytokines. Two biologically independent experiments for clone c11, one for clone c17 and c19. **c**, Quantitative transcript analysis of mRNAs encoding PGC transcription factors during differentiation without PGCLC cytokines at indicated time point. Expression levels are normalized to TBP; values are means ± s.d.; $n = 3$ biologically independent replicates. **d**, Whole-mount immunostaining of aggregates of *Otx2*$^{lacZ/GFP}$ cells at day 2 in the absence of cytokines for OTX2, BLIMP1 and AP2γ; scale bar; 40 μm; $n = 3$.

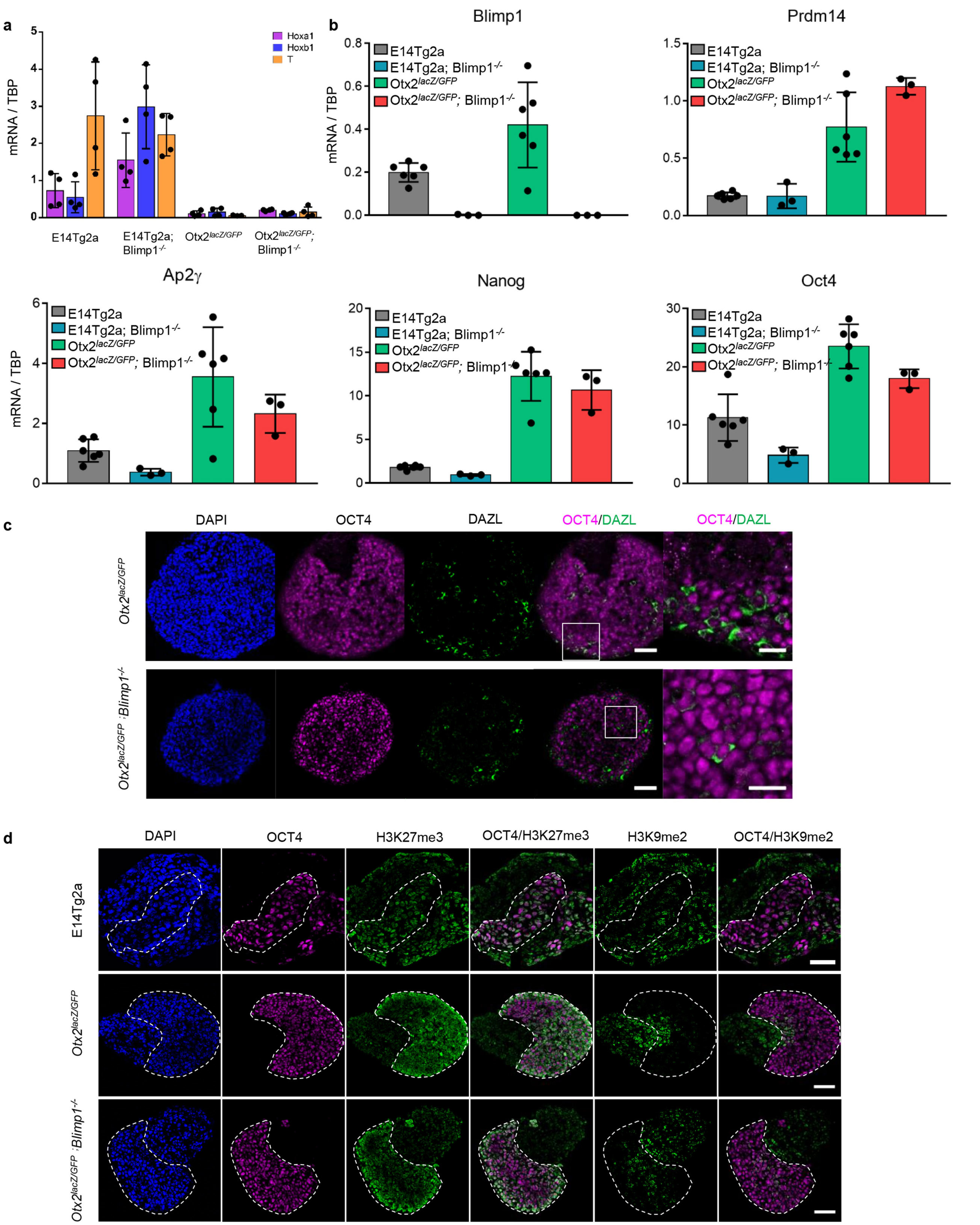




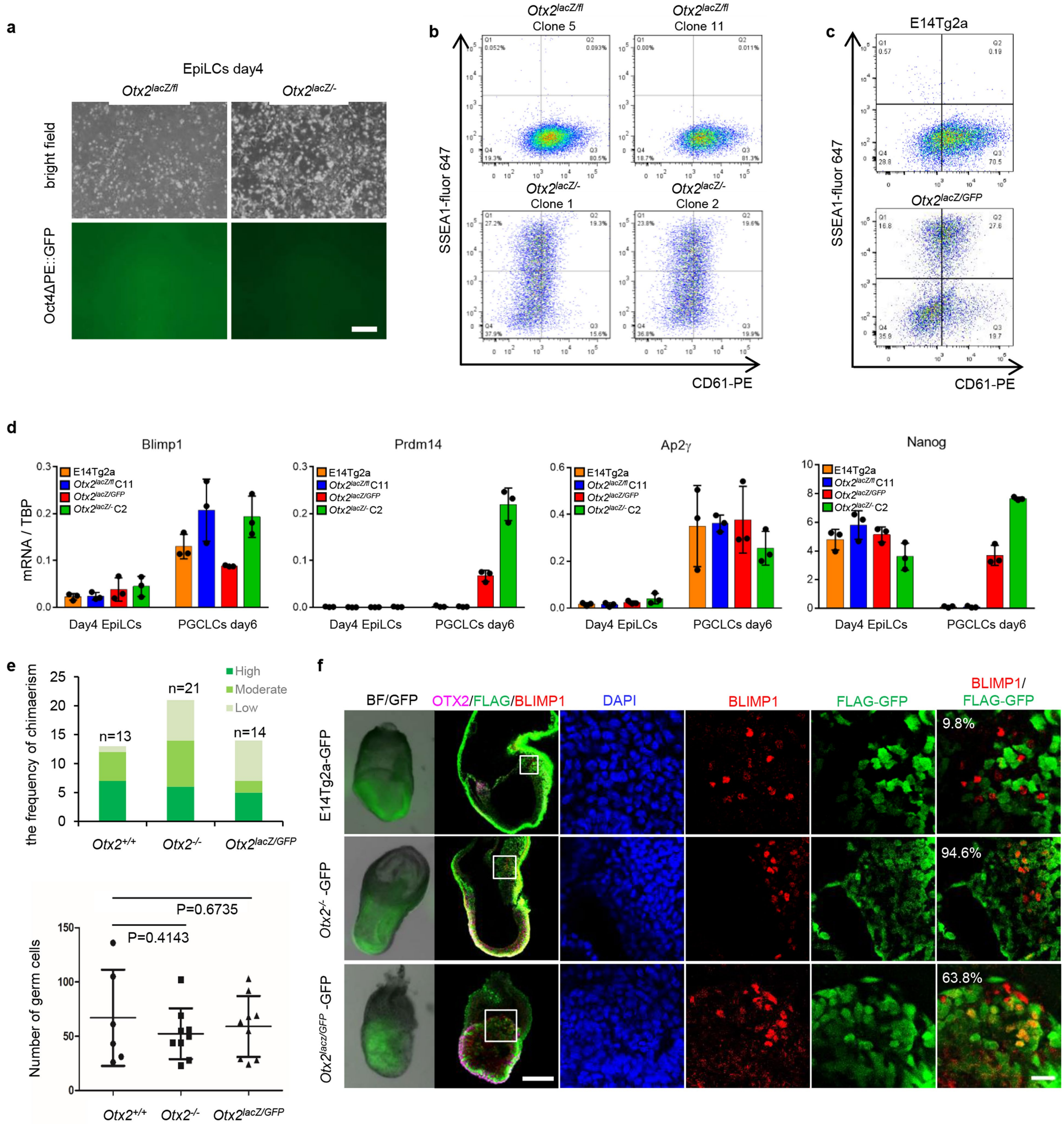
Extended Data Fig. 6 | Transcriptome analysis of EpiLCs and day 6 PGCLCs. a, **b**, Heat map of the normalized gene expression and principal component analysis of microarray data (from three biologically independent replicates under seven different conditions) ordered by unsupervised hierarchical clustering; rows correspond to transcripts and columns to cells. Differentiations performed in the presence (+Cyt) or absence (-Cyt) of cytokines are indicated. WT, E14Tg2a; $O^{-/-}$, *Otx2*$^{lacZ/GFP}$; $O^{-/-}$ $B^{-/-}$, *Otx2*$^{lacZ/GFP}$; *Blimp1*$^{-/-}$.



Extended Data Fig. 7 | PGCLC induction of independent *Blimp1*-null cell lines. a, Scheme showing the strategy used to generate *Blimp1*-knockout cell lines. A pair of gRNAs flanking *Blimp1* exon5 were co-expressed to ensure complete deletion of *Blimp1* exon5. Red arrows represent genotyping primer pairs used to screen clones. **b**, **c**, *Blimp1*-null clones used in Fig. 3 (**b**) or Extended Data Fig. 8d (**c**) were genotyped using primers indicated in **a**; $n = 2$ biologically independent replicates for both, all clones have been sequenced. **d**, Cells of the indicated genotypes (**c**) were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of aggregation in the presence of PGC induction cytokines; $n = 2$. **e**, **f**, Cells of the indicated genotypes (**c**) were assessed by flow cytometry for surface expression of SSEA1 and CD61 at day 6 of aggregation in the absence of PGC induction cytokines; $n = 2$.



Extended Data Fig. 8 | *Otx2*$^{-/-}$*Blimp1*$^{-/-}$ PGCLCs activate PGC markers. a, Quantitative analysis of somatic transcripts at day 2 of PGCLC induction in the indicated cell lines. Expression levels are normalized to TBP; values are means ± s.d.; $n = 4$ biological replicates, each dot represents the value from one experiment. **b**, Quantitative analysis of PGC transcription factor transcripts at day 2 of PGCLC induction in the indicated cell lines. Expression levels are normalized to TBP; values are means ± s.d.; $n = 6$ biological replicates for E14Tg2A and *Otx2*$^{lazZ/GFP}$, and 4 for Blimp1-knockout cell lines each dot represents the value from one experiment. **c**, Immunofluorescence staining for OCT4 and DAZL of cryo-sections of *Otx2*$^{lacZ/GFP}$ and *Otx2*$^{lacZ/GFP}$*Blimp1*$^{-/-}$ aggregates at day 6 of PGCLC induction; scale bar, 50 μm and 20 μm; $n = 2$ biologically independent replicates. **d**, OCT4, H3K27me3 and H3K9me2 immunofluorescence analysis of cryo-sections of E14Tg2a, *Otx2*$^{lacZ/GFP}$ and *Otx2*$^{lacZ/GFP}$*Blimp1*$^{-/-}$ aggregates at day 6 of PGCLC induction; scale bar, 50 μm; $n = 2$ biologically independent replicates.

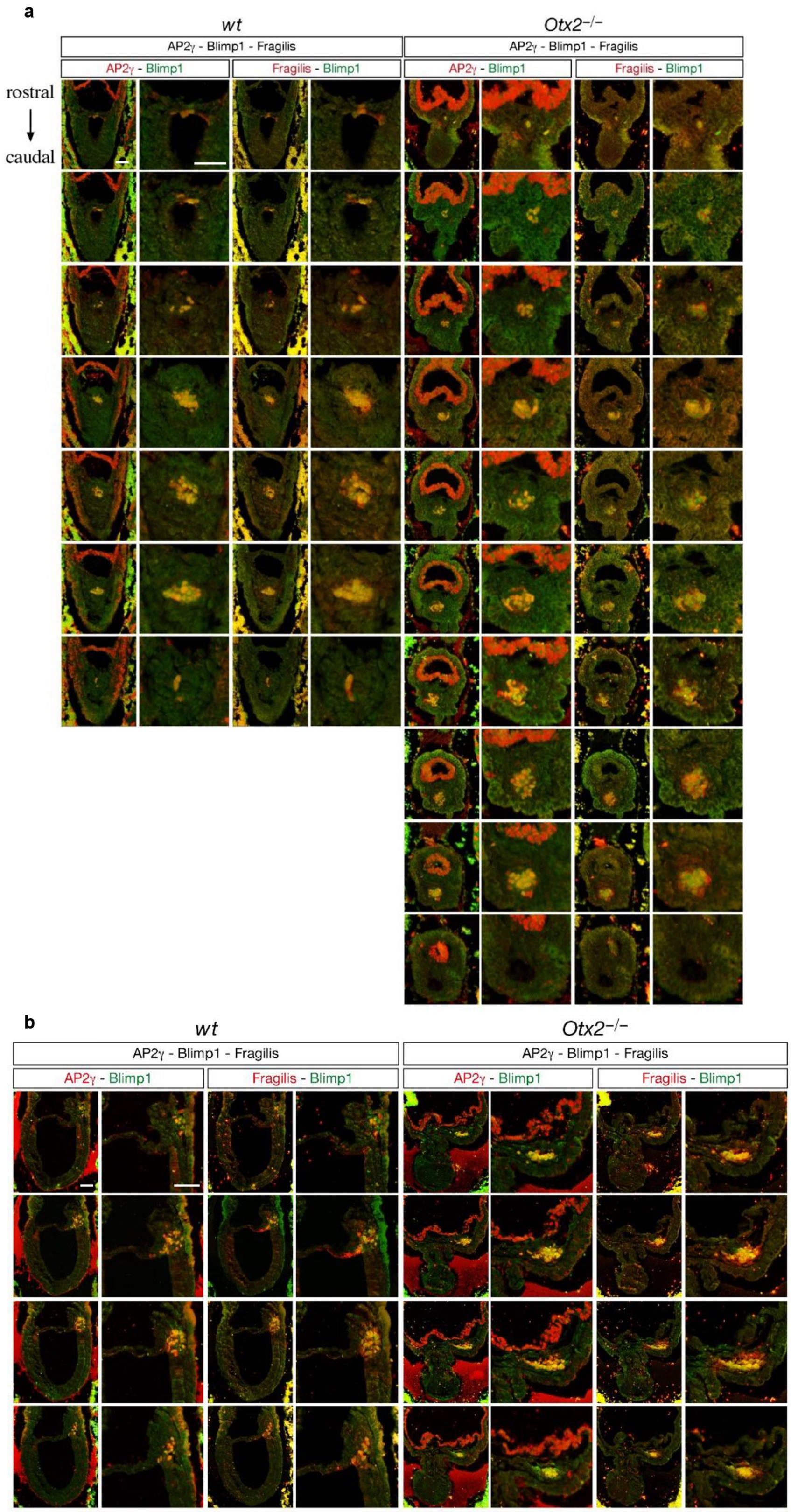


g

Cell lines	Embryos Transferred	No. of Embryos	No. of Chimaeras (% embryos)	No. for analysis
E14Tg2a-GFP	68	14	13 (92.8%)	6
Otx2$^{-/-}$ -GFP	86	33	21 (63.6%)	9
Otx2$^{lacZ/GFP}$ -GFP	57	18	14 (77.8%)	9

Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | OTX2 safeguards somatic lineages.
a. Representative morphologies and Oct4ΔPE::GFP expression of EpiSCs after one passage from EpiLCs ($n = 3$ for 1 clone of each genotype); scale bar; 200 μm. **b**, **c**, Flow cytometry analysis for surface expression of SSEA1 and CD61 at day 6 of PGCLC differentiation, initiated from EpiSCs after one passage from EpiLCs. One experiment for c5 and c1 and 6 biologically independent replicates for C11 and C2 (**b**); $n = 6$ biologically independent replicates (**c**). **d**, Quantitative transcript analysis of PGC transcription factors in the indicated cell lines. Expression levels are normalized to TBP; values are means ± s.d.; $n = 3$ biologically independent replicates, each dot represents the value from one experiment. **e**, Comparison of the frequency of degree of chimaerism (top) and the germ cell numbers (bottom, centre lines and error bars represents means ± s.d.) in E7.5 chimaeric embryos formed using wild-type or Otx2-null ESCs. *P* value (two-tailed unpaired *t*-test, 0.95 confidence intervals) is indicated. High, >70%; moderate, 30–70%; low, <30%. **f**, Bright-field and representative images of E7.5 chimaeric embryos formed by wild-type host embryos and GFP-labelled *Otx2*$^{+/+}$ ($n = 6$), *Otx2*$^{-/-}$ ($n = 9$) or *Otx2*$^{lacZ/GFP}$ ($n = 9$) ESCs assessed for GFP and BLIMP1/SOX2 expression, with magnified images of the proximal posterior regions. The proportion of BLIMP1-positive cells expressing GFP in the embryos is indicated; scale bar, 100 μm (left), 20 μm. **g**, Summary of embryo aggregations.



Extended Data Fig. 10 | *Otx2*-null embryos exhibit increased number of PGCs. a, **b**, Frontal-coronal (**a**) and sagittal (**b**) sections of wild-type and *Otx2*$^{-/-}$ E7.5 embryos stained with BLIMP1, AP2γ and Fragilis to detect PGCs. All sequential sections spanning the PGCLCs niche are shown; scale bar, 50 μm. The experiments were repeated in four wild-type and three *Otx2*-knockout embryos.

LETTER

https://doi.org/10.1038/s41586-018-0622-0

A metabolite-derived protein modification integrates glycolysis with KEAP1–NRF2 signalling

Michael J. Bollong[1,6], Gihoon Lee[2,3,6], John S. Coukos[2,3], Hwayoung Yun[1,5], Claudio Zambaldo[1], Jae Won Chang[2,3], Emily N. Chin[1], Insha Ahmad[1], Arnab K. Chatterjee[4], Luke L. Lairson[1,4]*, Peter G. Schultz[1,4]* & Raymond E. Moellering[2,3]*

Mechanisms that integrate the metabolic state of a cell with regulatory pathways are necessary to maintain cellular homeostasis. Endogenous, intrinsically reactive metabolites can form functional, covalent modifications on proteins without the aid of enzymes[1,2], and regulate cellular functions such as metabolism[3–5] and transcription[6]. An important 'sensor' protein that captures specific metabolic information and transforms it into an appropriate response is KEAP1, which contains reactive cysteine residues that collectively act as an electrophile sensor tuned to respond to reactive species resulting from endogenous and xenobiotic molecules. Covalent modification of KEAP1 results in reduced ubiquitination and the accumulation of NRF2[7,8], which then initiates the transcription of cytoprotective genes at antioxidant-response element loci. Here we identify a small-molecule inhibitor of the glycolytic enzyme PGK1, and reveal a direct link between glycolysis and NRF2 signalling. Inhibition of PGK1 results in accumulation of the reactive metabolite methylglyoxal, which selectively modifies KEAP1 to form a methylimidazole crosslink between proximal cysteine and arginine residues (MICA). This posttranslational modification results in the dimerization of KEAP1, the accumulation of NRF2 and activation of the NRF2 transcriptional program. These results demonstrate the existence of direct inter-pathway communication between glycolysis and the KEAP1–NRF2 transcriptional axis, provide insight into the metabolic regulation of the cellular stress response, and suggest a therapeutic strategy for controlling the cytoprotective antioxidant response in several human diseases.

In line with its role in responding to altered cellular conditions, numerous studies have linked deregulated KEAP1–NRF2 signalling in disease, including cancer[9], neurodegenerative disorders[10], chronic inflammatory diseases[11], diabetes[12] and ageing[13]. Efforts to target NRF2 signalling therapeutically have largely focused on covalent small-molecule agonists of KEAP1, including the clinical candidate bardoxolone methyl (BARD)[14]. To discover noncovalent modulators of the KEAP1–NRF2 signalling axis, as well as potentially new mechanisms of action for its regulation, we performed a cell-based, high-throughput phenotypic screen using a NRF2-dependent luciferase reporter (pTI-ARE-LUC) in IMR32 cells[15]. From a library of diverse heterocyclic compounds, we identified a hit compound, CBR-470-0, that did not contain any obvious electrophilic groups, and induced the

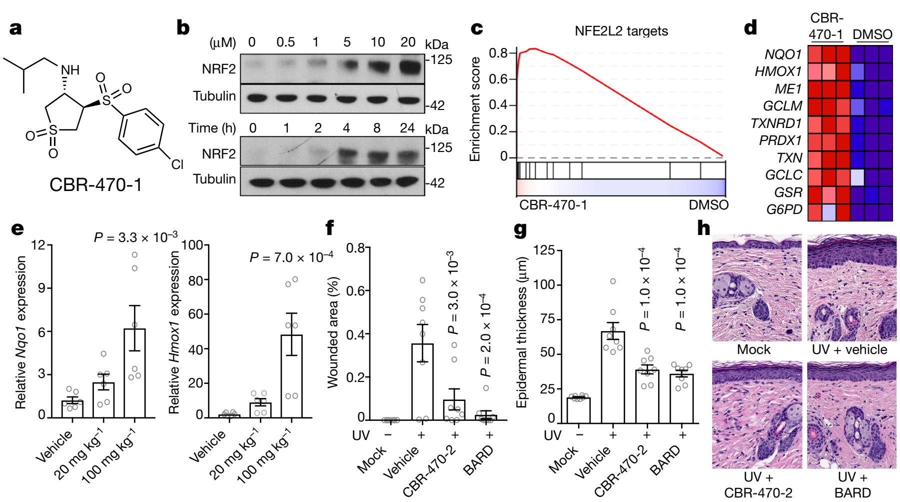


Fig. 1 | CBR-470-series compounds activate NRF2 signalling in vitro and in vivo. a, Structure of CBR-470-1. **b**, NRF2 protein levels from IMR32 cells treated with the indicated concentrations of CBR-470-1 for 4 h (top) or 5 µM CBR-470-1 for the indicated time periods (bottom). Blots are representative of three independent experiments. **c**, Gene set enrichment analysis (GSEA) plot depicting the enrichment of a NRF2 target gene set ('Singh_NFE2L2_Targets' in MSigDB) from IMR32 cells treated for 24 h with 5 µM CBR-470-1 ($n = 3$, $P < 0.0001$, nominal P value in GSEA). **d**, Heat map representation of the leading-edge subset of the most upregulated NRF2-regulated transcripts after CBR-470-1 treatment. Data are biologically independent samples. **e**, Relative *Nqo1* and *Hmox1* transcript levels in mouse skin tissue 24 h after the indicated oral doses of CBR-470-2 ($n = 6$, biologically independent samples). **f**, Quantification of wounded area by automated image analysis from animals of the indicated treatment groups at study end (day 10). **g**, Quantification of epidermal thickness from haematoxylin and eosin (H&E)-stained sections from the indicated groups at study end. **h**, Representative images of H&E-stained skin sections from animals euthanized at day 10 of the study. CBR-470-2, 50 mg kg^{-1} twice a day, orally. BARD, 3 mg kg^{-1} twice a day, orally. UV, 200 mJ cm^{-2}. Data are mean and s.e.m., $n = 8$ animals. Statistical analyses are by one-way analysis of variance (ANOVA) with Dunnett's correction (**e**–**g**).

[1]Department of Chemistry, The Scripps Research Institute, La Jolla, CA, USA. [2]Department of Chemistry, University of Chicago, Chicago, IL, USA. [3]Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL, USA. [4]California Institute for Biomedical Research (Calibr), La Jolla, CA, USA. [5]Present address: College of Pharmacy, Pusan National University, Busan, South Korea. [6]These authors contributed equally: Michael J. Bollong, Gihoon Lee. *e-mail: llairson@scripps.edu; schultz@scripps.edu; rmoellering@uchicago.edu

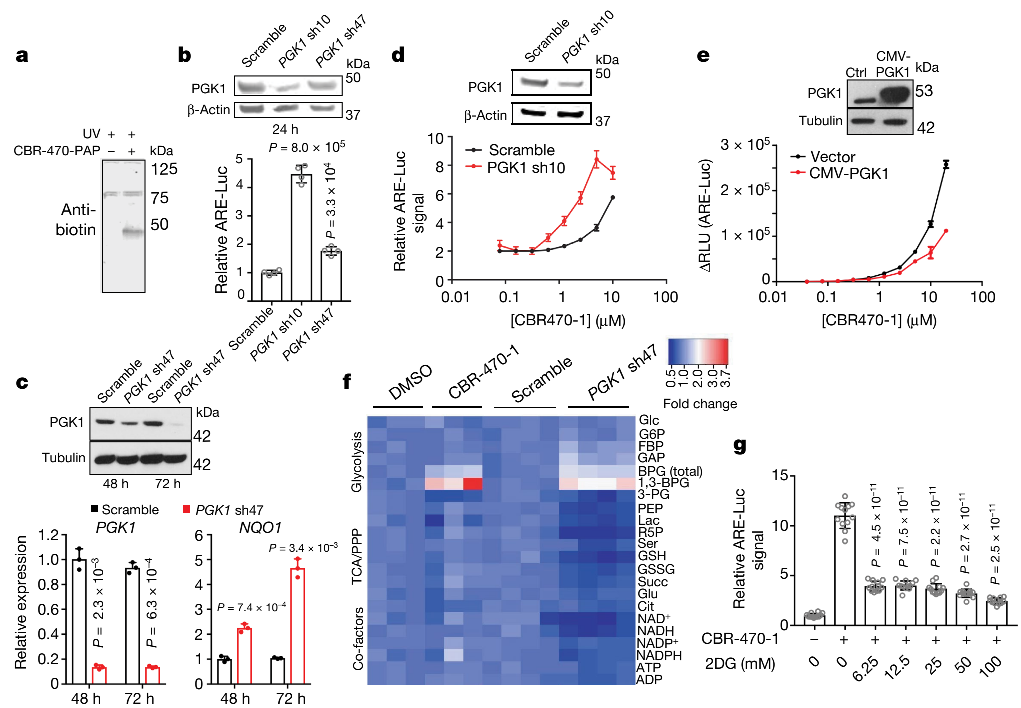


Fig. 2 | CBR-470-1-dependent inhibition of glycolysis activates NRF2 signalling. a, Anti-biotin western blot analysis of IMR32 cells treated with CBR-470-PAP (10 μM) for 1 h and exposed to UV light to induce photocrosslinking (representative shown from $n = 4$ biological replicates). **b**, Transient transfection of shRNA constructs targeting *PGK1* (sh10 and sh47) in HEK293T cells activates the ARE-LUC reporter. PGK1 and β-actin protein levels shown from representative experiments ($n = 4$ biological replicates). **c**, Viral shRNA knockdown of *PGK1* induces *NQO1* mRNA levels in IMR32 cells. PGK1 and tubulin protein levels are shown from representative experiments ($n = 3$). **d**, **e**, CBR-470-1 activation of ARE-LUC reporter in HEK293T cells with transient knockdown (**d**) or overexpression (**e**) of *PGK1* demonstrates opposing effects on compound potency. PGK1, actin and tubulin protein levels are shown from representative experiments ($n = 3$). ΔRLU, change in relative light units. **f**, Heat map depiction of relative metabolite levels in IMR32 cells treated for 30 min with CBR-470-1 (left) or viral shRNA knockdown of *PGK1* (right) relative to DMSO and scramble shRNA controls, respectively. BPG refers to both 2,3-BPG and 1,3-BPG, whereas 1,3-BPG refers specifically to the 1,3-isomer. 3-PG; 3-phosphoglycerate; Cit, citrate; FBP, fructose-1,6-bisphosphate; G6P, glucose 6-phosphate; GAP, glyceraldehyde-3-phosphate; Glc, glucose; Glu, glutamate; GSSG, glutathione disulfide; Lac, lactate; $NADP^+$, nicotinamide adenine dinucleotide phosphate; NAD^+, nicotinamide adenine dinucleotide; NADH, nicotinamide adenine dinucleotide (reduced); NADPH, nicotinamide adenine dinucleotide phosphate (reduced); PEP, 2-phosphoenolpyruvate; PPP, pentose phosphate pathway; R5P, ribose-5-phosphate; Succ, succinate; TCA, tricarboxylic acid. **g**, ARE-LUC reporter activity in IMR32 cells co-treated with CBR-470-1 (5 μM) and 2-deoxyglucose (2DG) for 24 h ($n = 12$). Statistical analyses are by univariate two-sided *t*-tests (**b**, **c**, **g**). Data are mean and s.d. of biologically independent samples.

transcriptional activity of antioxidant-response elements (AREs) to a similar magnitude to the previously reported NRF2 activators tert-butylhydroquinone (TBHQ) and AI-1 (Extended Data Fig. 1a–c). Structure–activity relationship analysis around the cyclic sulfone scaffold led to the identification of CBR-470-1 (Fig. 1a), an isobutylamine-substituted analogue that was not reactive in glutathione assays and had a half-maximum effective concentration (EC_{50}) value of approximately 1 μM in the cellular ARE-LUC assay (Extended Data Fig. 1d, e). CBR-470-1 treatment resulted in a dose- and time-dependent accumulation of NRF2 protein in IMR32 cells (Fig. 1b), and increased both mRNA and protein levels of the NRF2-responsive genes *NQO1* and *HMOX1* (Extended Data Fig. 1f, g; Extended Data Table 1). Expression profiling of IMR32 cells exposed to compound for 24 h revealed that the most significantly enriched gene set was 'NFE2L2 targets', which consisted of NRF2 target genes (Fig. 1c, d); the expression changes of these target transcripts were validated by focused quantitative reverse transcription PCR (qRT–PCR) analysis (Extended Data Fig. 1f). CBR-470-1 also induced transcript levels of *NQO1* and *HMOX1*, and enhanced NRF2 protein stabilization in HEK293T, SH-SY5Y and primary human lung fibroblasts (Extended Data Fig. 1h, i), indicating that these effects are not restricted to a specific cell type. Genetic depletion of NRF2 protein using short hairpin RNA (shRNA) inhibited the ability of CBR-470-1 and TBHQ to induce luciferase expression, indicating that CBR-470-1 activity is dependent on NRF2 (Extended Data Fig. 1j). Finally, CBR-470-1 treatment induced a cytoprotective NRF2 phenotype in vitro, as shown by the protection of SH-SY5Y cells challenged with the cell-permeable peroxide tert-butyl hydroperoxide (TBHP) (Extended Data Fig. 1k).

We next sought to determine whether CBR-470-1 or related analogues induce the activation of NRF2 signalling in vivo. Oral dosing of BALB/c mice with 20 or 100 mg kg^{-1} (twice a day, orally) of a glycine-substituted analogue, CBR-470-2, which is equally potent to CBR-470-1 and has more favourable bioavailability (Extended Data Fig. 2a–e), induced NRF2 target gene expression in several organs, with dose-dependent increases in *Nqo1* and *Hmox1* transcript levels observed in the skin (Fig. 1e). Published studies in *Nrf2* (also known as *Nfe2l2*)-knockout mice have demonstrated that NRF2 is essential to protect against photo-ageing phenotypes and skin carcinogenesis[16] resulting from ultraviolet (UV) irradiation[17]. We therefore evaluated CBR-470-2 activity in this acute UV damage mouse model. Mice were prophylactically dosed with CBR-470-2 (50 mg kg^{-1}, twice a day, orally) or BARD (3 mg kg^{-1}, orally) for five days before exposure to a single dose of UV irradiation. Mice were then dosed for an additional five days, euthanized and UV damage was quantified. CBR-470-2 and BARD treatment resulted in comparable beneficial effects on erythema histological scores and total wounded area (Fig. 1f, Extended Data Fig. 2f, g). Both CBR-470-2 and BARD were also found to decrease epidermal thickness in response to UV exposure, consistent with activation of the NRF2 cytoprotective program[18] (Fig. 1g, h). The combined pharmacodynamic and efficacy data indicate that CBR-470-2 treatment is capable of modulating NRF2 signalling in vivo, despite this compound series operating via an apparent mechanism that is independent of direct KEAP1 binding.

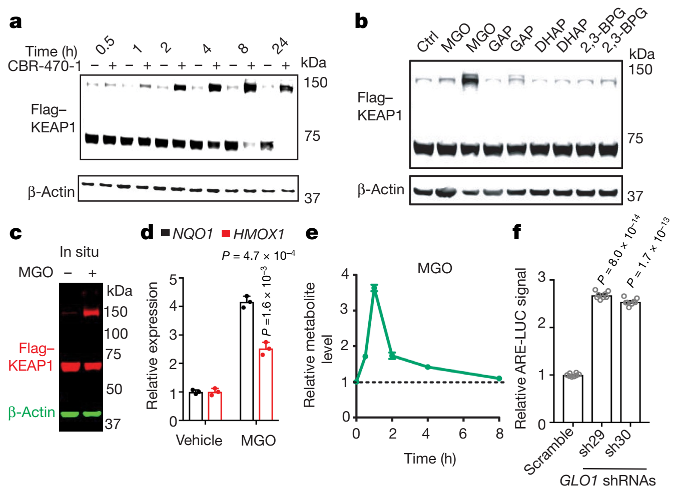


Fig. 3 | Methylglyoxal modifies KEAP1 to form a covalent, high-molecular mass dimer and activate NRF2 signalling. a, Time-course, anti-Flag western blot analysis of whole-cell lysates from HEK293T cells expressing Flag–KEAP1 treated with DMSO or CBR-470-1. **b**, Western blot monitoring of Flag–KEAP1 migration in HEK293T lysates after incubation with central glycolytic metabolites in vitro (1 and 5 mM, left and right for each metabolite). **c**, Flag–KEAP1 (red) and β-actin (green) from HEK293T cells treated with MGO (5 mM) for 8 h. **d**, Relative *NQO1* and *HMOX1* mRNA levels in IMR32 cells treated with MGO (1 mM) or water control ($n = 3$). **e**, LC–MS/MS quantitation of cellular MGO levels in IMR32 cells treated with CBR-470-1 relative to DMSO ($n = 4$). **f**, ARE-LUC reporter activity in HEK293T cells with transient shRNA knockdown of *GLO1* ($n = 8$). Statistical analyses are by univariate two-sided *t*-test (**d**, **f**). Data are mean ± s.e.m. of biologically independent samples.

To determine the mechanism by which CBR-470-1 activates NRF2 signalling, we generated a photo-affinity probe that contained biotin and diazirine substituents, termed CBR-470-PAP, which retained cellular activity in ARE-LUC reporter assays ($EC_{50} = 2.4\,\mu M$; Extended Data Fig. 3a, b). Treatment of IMR32 cells with 5 μM CBR-470-PAP for 1 h, followed by UV irradiation and subsequent anti-biotin western blot analysis of cellular lysates (Fig. 2a), together with biochemical fractionation and liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis, identified the enzyme phosphoglycerate kinase 1 (PGK1) as a potential target of CBR-470-PAP (Extended Data Fig. 3c). In vitro binding experiments with recombinant protein revealed that CBR-470-PAP selectively labelled PGK1, which was blocked with the soluble competitor CBR-470-1, or shRNA depletion of PGK1 protein levels (Extended Data Fig. 3d–f). Thermal stability assays in the presence of CBR-470-1 resulted in a consistent shift in PGK1 stability, and isothermal dose response profiling[19] against PGK1 and GAPDH also confirmed the selective, dose-dependent alteration of PGK1 stability in the presence of CBR-470-1 (Extended Data Fig. 3g–i). Furthermore, transient and viral shRNA knockdown of *PGK1* in IMR32 cells activated the ARE-LUC reporter signal and expression of *NQO1* (Fig. 2b, c). Knockdown or overexpression of PGK1 protein modulated the NRF2 reporter, with decreased or increased observed CBR-470-1 EC_{50} values, respectively (Fig. 2d, e). Finally, depletion of enolase 1, an enzyme downstream of PGK1, was also found to induce the ARE-LUC signal in IMR32 cells (Extended Data Fig. 3j). These results suggest that CBR-470-1 modulation of PGK1 activity, and therefore glycolysis, regulates NRF2 activation.

Consistent with the PGK1 inhibitory activity of CBR-470-1 (Extended Data Fig. 4a, b), targeted metabolomic profiling[4,20] of IMR32 cells treated with compound revealed a rapid increase in metabolite levels upstream of PGK1 (1,3- and 2,3-bisphosphoglycerate (BPG), and D-glyceraldehyde-3-phosphate (GAP)), and depletion of downstream metabolites such as 3-phosphoglycerate and lactate, which mirrored the profile observed upon viral knockdown of *PGK1* in IMR32 cells (Fig. 2f, Extended Data Fig. 4c, d, Extended Data Table 2). In addition, the co-treatment of ARE-LUC-expressing reporter cells with CBR-470-1 and an inhibitor of glucose entry into glycolysis, 2-deoxyglucose, significantly reduced reporter activation in a dose-dependent manner (Fig. 2g). Together, these data suggested that glycolytic intermediates may serve as a signal to the NRF2 signalling axis.

We first investigated whether 1,3-BPG, which is directly metabolized by PGK1, could be involved in signalling to the KEAP1–NRF2 pathway via phosphoglyceryl-lysine modification of KEAP1. However, CBR-470-1 treatment of IMR32 cells for 30 min, a time at which 1,3-BPG levels are increased, did not result in altered KEAP1 levels, or any anti-phosphoglyceryl-lysine immunoreactive bands using polyclonal antibodies raised against the phosphoglyceryl-lysine epitope (Extended Data Fig. 5a–c). These western blots did, however, reveal the appearance of a CBR-470-1 dose-dependent, high-molecular mass KEAP1 (HMM-KEAP1) band at roughly twice the molecular mass of monomeric KEAP1 (Fig. 3a). The HMM-KEAP1 band was stable to reduction (Extended Data Fig. 5d) and exhibited kinetics and dose-dependent formation consistent with CBR-470-1-dependent NRF2 stabilization and *NQO1* induction (Fig. 1b), but distinct from the direct KEAP1 alkylator TBHQ (Extended Data Fig. 5e–j). Co-treatment of cells with CBR-470-1 and either reduced glutathione (GSH) or *N*-acetylcysteine inhibited HMM-KEAP1 band formation (Extended Data Fig. 5k). Knockdown of *PGK1*, which activates NRF2 target gene expression, also formed HMM-KEAP1, and this band was competed by co-treatment with GSH (Extended Data Fig. 5l). Together, these data indicated that modulation of glycolysis by CBR-470-1 results in the formation of HMM-KEAP1 that is consistent with a covalent KEAP1 dimer, which has been previously observed[21–23], but remained uncharacterized at the molecular level.

Several central glycolytic metabolites other than 1,3-BPG contain reactive functionalities, including the triosephosphate isomers GAP and dihydroxyacetone phosphate, as well as their non-enzymatic elimination product methylglyoxal (MGO), an electrophilic dicarbonyl compound that has been found to form numerous modifications on nucleophilic residues in proteins[24,25]. Among these candidates, only treatment of cell lysates or live cells with MGO resulted in the selective formation of HMM-KEAP1 (Fig. 3b, c). Treatment of Flag–KEAP1-containing lysates or purified KEAP1 with freshly distilled MGO induced dose-dependent formation of HMM-KEAP1 at mid-micromolar concentrations (Extended Data Fig. 5m, n), which is consistent with the range of MGO concentrations previously reported in living cells[26,27]. MGO treatment in cells functionally activated expression of the downstream NRF2 target genes *NQO1* and *HMOX1* (Fig. 3d). Targeted LC–MS measurement of derivatized MGO confirmed that CBR-470-1 treatment resulted in a significant increase in cellular MGO levels in the first few hours of treatment (Fig. 3e, Extended Data Fig. 6a–c), which, along with pathway activation, was sensitive to GSH treatment (Extended Data Fig. 6d–f). To test the involvement of MGO in KEAP1–NRF2 signalling further, we perturbed its degradation, which is mediated by GSH and glyoxylase 1 (GLO1). Knockdown of *GLO1* by shRNA resulted in ARE-LUC reporter activation (Fig. 3f), and also sensitized cells to CBR-470-1 activation of the ARE-LUC reporter (Extended Data Fig. 6g). Direct modulation of GLO1 enzymatic activity with a cell-permeable inhibitor (GLOi) also amplified reporter activation by CBR-470-1 (Extended Data Fig. 6h). Collectively, these metabolomic, proteomic and transcriptomic data established shared kinetics between MGO accumulation, HMM-KEAP1 formation and NRF2 pathway activation, suggesting the existence of a direct link between glycolysis and the KEAP1–NRF2 signalling pathway that is mediated by the modification of KEAP1 by MGO.

A stable isotope labelling with amino acids in cell culture (SILAC)-based quantitative proteomic approach (Extended Data Fig. 7a) suggested the N-terminal region (amino acids 1–50) and BTB domains (amino acids 150–169) as candidate domains and residues that could be involved in HMM-KEAP1 formation in response to the CBR-470-1-induced increase in MGO (Extended Data Fig. 7b–d). We therefore examined more than a dozen C-to-S, K-to-M/R and R-to-A mutations

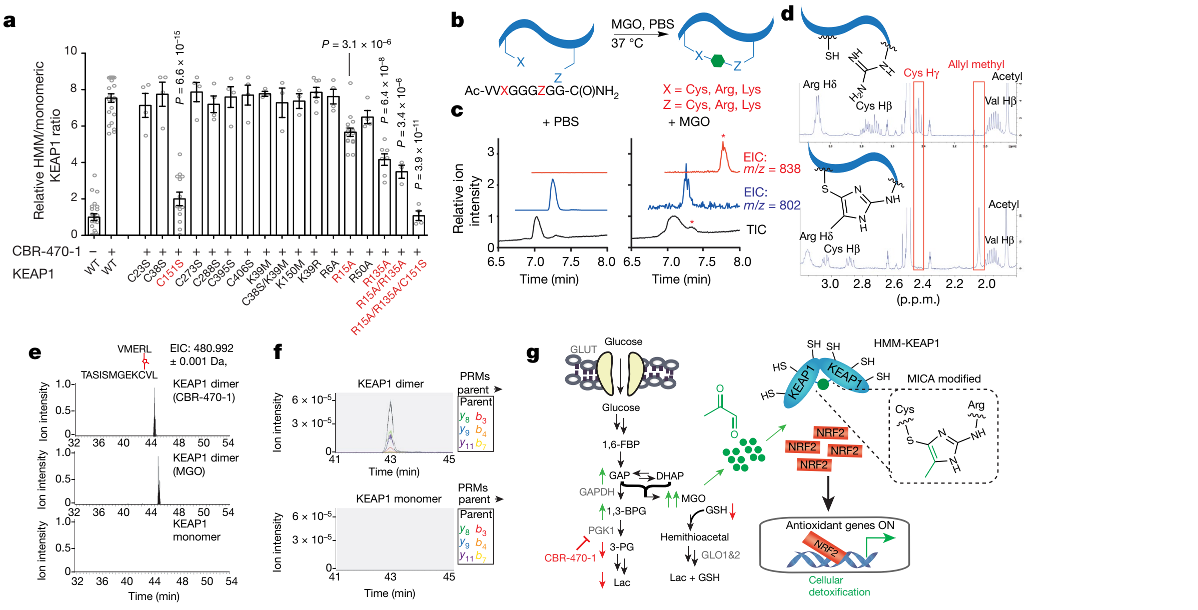


Fig. 4 | Methylglyoxal forms a posttranslational modification between proximal cysteine and arginine residues in KEAP1. a, Quantified HMM-KEAP1 formation of wild-type (WT) or mutant Flag–KEAP1 from HEK293T cells treated with DMSO or CBR-470-1 for 8 h ($n = 23$ for WT; $n = 16$ for R15A; $n = 13$ for C151S; $n = 7$ for K39R, R135A; $n = 4$ for R6A, R50A, all other C-to-S mutations, and R15/135A and C151S triple-mutant; $n = 3$ for R15/135A, and all K-to-M mutations). **b**, Schematic of the model peptide screen for intramolecular modifications formed by MGO and nucleophilic residues. **c**, Total ion and extracted ion chromatograms (TICs and EICs, respectively) from MGO- and mock-treated peptide, with a new peak in the former condition marked by an asterisk. EICs are specific to the indicated m/z ($n = 3$ independent biological replicates). **d**, ^{1}H-NMR spectra of the unmodified (top) and MICA-modified (bottom) model peptide, with pertinent protons highlighted in each. Notable changes in the MICA-modified spectrum include the appearance of a singlet at 2.04 p.p.m., loss of the thiol proton at 2.43 p.p.m., and changes in chemical shift and splitting pattern of the cysteine beta protons and the arginine delta and epsilon protons. Full spectra and additional multidimensional NMR spectra can be found in Extended Data Fig. 7. **e**, EIC from LC–MS/MS analyses of gel-isolated and digested HMM-KEAP1 (CBR-470-1 and MGO-induced) and monomeric KEAP1 for the C151–R135 crosslinked peptide. Slight retention time variation was observed on commercial columns ($n = 3$ independent biological replicates). **f**, PRM chromatograms for the parent and six parent-to-daughter transitions in representative targeted proteomic runs from HMM-KEAP1 and monomeric digests ($n = 6$). **g**, Schematic depicting the direct communication between glucose metabolism and KEAP1–NRF2 signalling mediated by MGO modification of KEAP1 and subsequent activation of the NRF2 transcriptional program. Statistical analyses are by univariate two-sided t-test (**a**). Data are mean ± s.e.m. of biologically independent samples.

within these domains, as well as other known functional residues in KEAP1, for their effect on HMM-KEAP1 formation. Two arginine residues (R15 of the N-terminal region domain and R135 of the BTB domain) significantly, but incompletely, reduced the formation of HMM-KEAP1 (Fig. 4a). More notable was the near complete inhibition of HMM-KEAP1 formation of the C151S mutant in the BTB domain (Fig. 4a). Consistent with this effect, levels of C151-containing tryptic peptides were reduced by MGO treatment, and the pre-treatment of cells with BARD, which alkylates C151, inhibited HMM-KEAP1 formation (Extended Data Fig. 7d, e). C151 lies in an exposed region of the BTB domain that is predicted to mediate the homodimeric interface between two KEAP1 monomers, which is necessary for proper NRF2 binding and ubiquitination[8,23]. Therefore, the strong abrogation of HMM-KEAP1 formation by the mutation of C151 and proximal arginine residues suggested that MGO may be mediating an uncharacterized modification between these residues.

To identify this modification, we synthesized a model peptide containing cysteine and arginine separated by a glycine linker, which was intended to mimic high inter- or intramolecular cysteine and arginine proximity, and treated it with MGO at physiological temperature and pH overnight (Fig. 4b). LC–MS analysis revealed a new peak, which corresponded to a mass increase of 36 Da, consistent with a mercaptomethylimidazole crosslink (Fig. 4c, d) formed by nucleophilic attack of the dicarbonyl by the side chains of cysteine and arginine, followed by dehydration-mediated formation of a methylimidazole crosslink between cysteine and arginine (MICA) posttranslational modification. We purified this product and confirmed its structure by a series of one- and two-dimensional NMR experiments (Fig. 4d, Extended Data Fig. 7f–i). To determine whether MICA modification occurs within KEAP1 protein, we treated cells with CBR-470-1 or MGO, isolated HMM-KEAP1 and monomeric KEAP1 by gel electrophoresis, and then digested these discrete populations for LC–MS/MS analyses. A peptide bearing a MICA crosslink between C151 and R135 was identified in isolated HMM-KEAP1 after both CBR-470-1 and MGO treatment, but not in the isolated monomeric KEAP1 (Fig. 4e). Furthermore, parallel-reaction monitoring (PRM) confirmed the presence and co-elution of more than a dozen parent-to-daughter ion transitions that were uniquely present in HMM-KEAP1 (Fig. 4f, Extended Data Fig. 8a, b). These studies suggest a model in which glycolytic metabolic status is coupled to NRF2-dependent gene expression through the direct interaction of a reactive glycolytic metabolite, MGO, and the sentinel protein KEAP1 via the formation of a stable and mechanistically novel protein posttranslational modification (Fig. 4g).

Although MGO has been previously shown to form covalent modifications on diverse proteins, the compositions, sites and functions of these modifications have remained largely unknown. Similarly, several recent reports have implicated MGO in the pathogenesis of diseases such as diabetes[28] and ageing[29], but the discrete molecular targets of MGO in these contexts are unidentified. Here we found that inhibition of PGK1 increases levels of triosephosphates, which results in increased levels of cellular MGO and the formation of a HMM-KEAP1 species that leads to NRF2-dependent gene expression. Formation of the HMM-KEAP1 species involves the posttranslational modification MICA that is dependent on MGO and forms a covalent linkage between proximal cysteine and arginine residues. These results raise intriguing questions about the general reactivity of MGO, its potential role as a signalling metabolite in other cellular processes, and the specific modifications involved in often-cited advanced glycated end products as biomarkers of disease pathology. We have shown that both cellular and lysate treatment with MGO results in the selective modification of

C151 in KEAP1, probably owing to the intrinsic hyperreactivity of this residue, and the presence of properly oriented arginine(s) that enables formation of the MICA modification. Additional factors such as local metabolite concentration gradients may also contribute to MICA formation in KEAP1. Future studies to determine the full target profile of MGO, and specifically other inter- and intramolecular MICA modifications, are expected to shed light on this model and provide a global view of MGO modification sites in the proteome.

The direct connection between glucose metabolism and the KEAP1–NRF2 axis by MGO adds a further layer of regulation to both pathways and global metabolic status. First, this connection highlights the role of glycolysis in regulating cellular redox status beyond the contribution to NADPH and glutathione production. These reducing equivalents are crucial for the regulation of a wide range of reactive species in the cell, and when these levels are deregulated, the KEAP1–NRF2 pathway is poised to respond and limit cellular damage. Recent studies have also implicated the output of the NRF2 transcriptional program in the direct detoxification of MGO through increased glutathione synthesis[30], *GLO1* transcription, as well as the redirection of glucose carbon away from central metabolites (for example, MGO) and into the pentose phosphate pathway[31]. Therefore, the direct coupling of glucose metabolism with KEAP1 function through MGO creates an intrinsic feedback loop to sense and respond to changing metabolic demands in the cell. A final aspect of this study is the notion that modulation of endogenous reactive metabolite levels using small molecules may represent an alternative approach towards activating the ARE pathway for the treatment of several diseases that involve metabolic stress.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0622-0.

Received: 22 October 2017; Accepted: 21 August 2018;
Published online 15 October 2018.

1. Lin, H., Su, X. & He, B. Protein lysine acylation and cysteine succination by intermediates of energy metabolism. *ACS Chem. Biol.* **7**, 947–960 (2012).
2. Wagner, G.R., et al. A class of reactive acyl-CoA species reveals the non-enzymatic origins of protein acylation. *Cell Metab.* **25**, 823–837 (2017).
3. Zhang, Z. et al. Identification of lysine succinylation as a new post-translational modification. *Nat. Chem. Biol.* **7**, 58–63 (2011).
4. Moellering, R. E. & Cravatt, B. F. Functional lysine modification by an intrinsically reactive primary glycolytic metabolite. *Science* **341**, 549–553 (2013).
5. Weinert, B. T. et al. Acetyl-phosphate is a critical determinant of lysine acetylation in *E. coli*. *Mol. Cell* **51**, 265–272 (2013).
6. Sabari, B. R., Zhang, D., Allis, C. D. & Zhao, Y. Metabolic regulation of gene expression through histone acylations. *Nat. Rev. Mol. Cell Biol.* **18**, 90–101 (2016).
7. Kobayashi, A. et al. Oxidative and electrophilic stresses activate Nrf2 through inhibition of ubiquitination activity of Keap1. *Mol. Cell. Biol.* **26**, 221–229 (2006).
8. Lo, S. C., Li, X., Henzl, M. T., Beamer, L. J. & Hannink, M. Structure of the Keap1:Nrf2 interface provides mechanistic insight into Nrf2 signaling. *EMBO J.* **25**, 3605–3617 (2006).
9. Jaramillo, M. C. & Zhang, D. D. The emerging role of the Nrf2-Keap1 signaling pathway in cancer. *Genes Dev.* **27**, 2179–2191 (2013).
10. Scannevin, R. H. et al. Fumarates promote cytoprotection of central nervous system cells against oxidative stress via the nuclear factor (erythroid-derived 2)-like 2 pathway. *J. Pharmacol. Exp. Ther.* **341**, 274–284 (2012).
11. Khor, T. O. et al. Nrf2-deficient mice have an increased susceptibility to dextran sulfate sodium-induced colitis. *Cancer Res.* **66**, 11580–11584 (2006).
12. Uruno, A. et al. The Keap1-Nrf2 system prevents onset of diabetes mellitus. *Mol. Cell. Biol.* **33**, 2996–3010 (2013).
13. Sykiotis, G. P. & Bohmann, D. Keap1/Nrf2 signaling regulates oxidative stress tolerance and lifespan in *Drosophila*. *Dev. Cell* **14**, 76–85 (2008).
14. Cleasby, A. et al. Structure of the BTB domain of Keap1 and its interaction with the triterpenoid antagonist CDDO. *PLoS One* **9**, e98896 (2014).
15. Hur, W. et al. A small-molecule inducer of the antioxidant response element. *Chem. Biol.* **17**, 537–547 (2010).
16. Saw, C. L. et al. Impact of Nrf2 on UVB-induced skin inflammation/photoprotection and photoprotective effect of sulforaphane. *Mol. Carcinog.* **50**, 479–486 (2011).
17. Tao, S., Justiniano, R., Zhang, D. D. & Wondrak, G. T. The Nrf2-inducers tanshinone I and dihydrotanshinone protect human skin cells and reconstructed human skin against solar simulated UV. *Redox Biol.* **1**, 532–541 (2013).
18. El-Abaseri, T. B., Putta, S. & Hansen, L. A. Ultraviolet irradiation induces keratinocyte proliferation and epidermal hyperplasia through the activation of the epidermal growth factor receptor. *Carcinogenesis* **27**, 225–231 (2006).
19. Martinez Molina, D. et al. Monitoring drug target engagement in cells and tissues using the cellular thermal shift assay. *Science* **341**, 84–87 (2013).
20. Chang, J. W., Lee, G., Coukos, J. S. & Moellering, R. E. Profiling reactive metabolites via chemical trapping and targeted mass spectrometry. *Anal. Chem.* **88**, 6658–6661 (2016).
21. Zhang, D. D. & Hannink, M. Distinct cysteine residues in Keap1 are required for Keap1-dependent ubiquitination of Nrf2 and for stabilization of Nrf2 by chemopreventive agents and oxidative stress. *Mol. Cell. Biol.* **23**, 8137–8151 (2003).
22. Wakabayashi, N. et al. Protection against electrophile and oxidant stress by induction of the phase 2 response: fate of cysteines of the Keap1 sensor modified by inducers. *Proc. Natl Acad. Sci. USA* **101**, 2040–2045 (2004).
23. Ogura, T. et al. Keap1 is a forked-stem dimer structure with two large spheres enclosing the intervening, double glycine repeat, and C-terminal domains. *Proc. Natl Acad. Sci. USA* **107**, 2842–2847 (2010).
24. Lo, T. W., Westwood, M. E., McLellan, A. C., Selwood, T. & Thornalley, P. J. Binding and modification of proteins by methylglyoxal under physiological conditions. A kinetic and mechanistic study with Nα-acetylarginine, Nα-acetylcysteine, and Nα-acetyllysine, and bovine serum albumin. *J. Biol. Chem.* **269**, 32299–32305 (1994).
25. Rabbani, N. & Thornalley, P. J. Dicarbonyl proteome and genome damage in metabolic and vascular disease. *Biochem. Soc. Trans.* **42**, 425–432 (2014).
26. Chaplen, F. W., Fahl, W. E. & Cameron, D. C. Evidence of high levels of methylglyoxal in cultured Chinese hamster ovary cells. *Proc. Natl Acad. Sci. USA* **95**, 5533–5538 (1998).
27. Dhar, A., Desai, K., Liu, J. & Wu, L. Methylglyoxal, protein binding and biological samples: are we getting the true measure? *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **877**, 1093–1100 (2009).
28. Moraru, A. et al. Elevated levels of the reactive metabolite methylglyoxal recapitulate progression of type 2 diabetes. *Cell Metab.* **27**, 926–934 (2018).
29. Ravichandran, M. et al. Impairing l-threonine catabolism promotes healthspan through methylglyoxal-mediated proteohormesis. *Cell Metab.* **27**, 914–925 (2018).
30. Nishimoto, S., Koike, S., Inoue, N., Suzuki, T. & Ogasawara, Y. Activation of Nrf2 attenuates carbonyl stress induced by methylglyoxal in human neuroblastoma cells: Increase in GSH levels is a critical event for the detoxification mechanism. *Biochem. Biophys. Res. Commun.* **483**, 874–879 (2017).
31. Mitsuishi, Y. et al. Nrf2 redirects glucose and glutamine into anabolic pathways in metabolic reprogramming. *Cancer Cell* **22**, 66–79 (2012).

Acknowledgements We thank S. Zhu for discussions about target identification experiments. Animal experiments were approved by the Scripps Research Institute Institutional Review Board. We are grateful for financial support of this work from the following: Kwanjeong Educational Fellowship (to G.L.); NIH MSTP Training Grant (T32GM007281 to J.S.C.); NIH R00CA175399, R01CA211916 and DP2GM128199 (R.E.M.); V Foundation for Cancer Research (V2015-020 to R.E.M.); Damon Runyon Cancer Research Foundation (DFS08-14); The Skaggs Institute for Chemical Biology, and The University of Chicago.

Reviewer information *Nature* thanks H. Christofk, A. Dinkova-Kostova, H. Lin and the anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors reviewed the manuscript. M.J.B., G.L., J.S.C., H.Y., C.Z., J.W.C., I.A., L.L.L. and R.E.M. designed and performed biochemical and cell-based biological experiments. H.Y., J.W.C., J.S.C. and A.K.C. synthesized and characterized chemical probes and reagents. E.N.C., M.J.B., L.L.L. and P.G.S. designed, performed and analysed in vivo experiments. G.L., J.S.C., J.W.C. and R.E.M. designed, performed and analysed metabolomic, proteomic and structural characterization experiments. L.L.L., P.G.S. and R.E.M. conceived of the study and supervised research. M.J.B. and R.E.M. wrote the manuscript with considerable input from all authors.

Competing interests Declared none.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0622-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0622-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to L.L.L., P.G.S. or R.E.M.

METHODS

No statistical methods were used to predetermine sample size. *Chemicals.* TBHQ, 2-deoxyglucose, MGO and GSH were obtained from Sigma Aldrich. The synthesis of AI-1 has been described previously[32]. The GLO1 inhibitor (CAS no. 174568-92-4) was from MedChemExpress. CBR-470-0 and CBR-581-9 were from ChemDiv. CBR-470-1 (initially from ChemDiv as D470-2172) and related analogues were synthesized in house according to full methods described in the Supplementary Information. All commercially obtained chemicals were dissolved in DMSO and used without further purification, with the exception of 2-deoxyglucose, MGO and GSH, which were delivered as aqueous solutions.

Cell culture. IMR32, SH-SY5Y, HeLa and HEK293T cell lines were purchased from ATCC. Human lung fibroblasts and mouse dermal fibroblasts (C57BL/6-derived) were obtained from Sciencell and used before passage 3. IMR32, human lung fibroblasts, SH-SY5Y, HeLa and HEK293T cells were propagated in DMEM (Corning) supplemented with 10% fetal bovine serum (FBS, Corning) and 1% penicillin/streptomycin (Gibco). Mouse dermal fibroblasts were propagated in fibroblast medium 2 from Sciencell. Mouse epidermal keratinocytes (MPEK-BL6) were obtained from Zen Bio and propagated in epidermal keratinocyte medium (Zen Bio).

High-throughput screening and ARE-LUC reporter assay. For high-throughput screening, IMR32 cells were plated at 5×10^3 cells per well in white 384-well plates in 40 μl of growth medium. The next day 100 ng of pTI-ARE-LUC reporter plasmid in 10 μl of Optimem medium (Gibco) was transfected into each well using Fugene HD at a dilution of 1 μg plasmid DNA:4 μl of Fugene. After 24 h, compounds were transferred using a 100 nl pintool head affixed to PerkinElmer FX instrument such that the final screening concentration was 2 μM. After 24 h incubation, ARE-LUC luminance values were recorded on an Envision instrument after the addition of 30 μl of Bright Glo reagent solution (Promega, diluted 1:3 in water). Compounds which increased ARE-LUC signal greater than four *Z*-scores from plate mean were deemed hits. For overexpression and knockdown experiments in HEK293T with ARE-LUC reporter readouts, 5×10^5 cells were plated on poly-D-lysine-coated plates and transfected with 1.5 μg of overexpression or shRNA plasmid and 500 ng of pTI-ARE-LUC using Optimem medium and Fugene in the same mode as above. After 24 h, 10^3 transfected cells were plated in 50 μl of growth medium in white 96-well plates. After a 24 h incubation, a further 50 μl of growth medium with compound at the indicated concentrated was added to each well. ARE-LUC luminance values were recorded on an Envision plate reader 24 h later by the addition of 75 μl of Bright Glo reagent solution (1:3 in water).

Peroxide stress model. Approximately 10^4 SH-SY5Y cells were plated in 100 μl of growth medium in white 96-well plates. After 48 h of compound treatment, 20 μl TBHP diluted to the indicated concentrations was added to each well. After an 8 h incubation, cell viability measurements were recorded on an Envision plate reader after the addition of 50 μl of a Cell Titer Glo solution (Promega, diluted 1:6 in water). Relative viabilities are reported as a fraction relative to the same dose of compound treatment without TBHP.

shRNA knockdown studies. PGK1-targeting shRNA vectors sh10 and sh47 refer to Sigma Mission shRNA lentiviral clones NM_000291.2-338s1c1 and NM_000291.2-935s1c1 respectively. GLO1-targeting shRNA vectors sh29 and sh30 refer to Sigma Mission shRNA lentiviral clones NM_006708.1-195s1c1 and NM_006708.1-292s1c1, respectively. The non-targeting scrambled control vector refers to SHC002 (Sigma). Lentiviruses were generated in HEK293T cells by transient expression of the above vectors with pSPAX2 and pMD2.G packaging vectors (Addgene plasmids 11260 and 12259). Viral supernatants were collected after 48 h of expression and passed through a 70-μm syringe filter before exposure to target cells.

qRT–PCR. Cells were collected by trypsinization and subsequent centrifugation at 500*g*. RNA was isolated using RNeasy kits from Qiagen and concentrations obtained using a NanoDrop instrument. Then 500 ng–5 μg of RNA was then reverse transcribed with oligo dT DNA primers using a SuperScript III First-Strand Synthesis kit from Invitrogen. Quantitative RT–PCR reactions were measured on a Viia 7 Real-Time PCR system (Thermo) using a Clontech SYBR green-based master mix. Gene-specific primer sets are provided in Extended Data Table 1. Reactions were normalized to *TUBG1* levels for each biological replicate and relative transcript abundance calculated using the comparative C_t method.

GSEA. Total RNA was extracted from IMR32 cells treated for 24 h with either DMSO or 5 μM CBR-470-1 (3 biological replicates per condition) using an RNeasy kit (Qiagen). RNA sequencing (RNA-seq) experiments were performed by the Scripps Next Generation Sequencing Core according to established in house methods. GSEAs and leading edge heat maps were generated with TPM values from the above experiment using the java GSEA package. 'NFE2L2 targets' gene set refers to Molecular Signature Database (http://software.broadinstitute.org/gsea/msigdb) gene set ID M2662.

Quantitative metabolomic profiling. For polar metabolite profiling experiments, cells were grown in 15 cm plates and cultured in RPMI supplemented with 10% FBS, 2 mM L-glutamine and 1% penicillin/streptomycin before media replacement containing either vehicle (DMSO) or the indicated dose of CBR-470-1. After incubation for the appropriate time, cells were scraped into ice-cold PBS and isolated by centrifugation at 1,400*g* at 4 °C. Cell pellets were resuspended in 300 μl of an 80:20 mixture of cold methanol/water, an internal standard was added (10 nmol d_3-serine; Sigma Aldrich), and the suspension was sonicated (Fisher Scientific, FB-505) for 5 s followed by a 10 min centrifugation at 16,000*g*. The supernatant was collected, dried under N_2 gas and resulting dried metabolites resuspended in 30 μl of 40% methanol/water for analysis on an Agilent triple quadrupole LC–MS/MS (Agilent Technologies, 6460 QQQ). For negative mode operation, metabolites were separated by hydrophilic interaction chromatography with a Luna-NH_2 column (Phenomenex) running mobile phase A (CH_3CN supplemented with 0.2% NH_4OH) and B (95:5 (v/v) H_2O:CH_3CN supplemented with 50 mM ammonium acetate and 0.2% NH_4OH) and the following gradient: 0% B for 3 min; linear increase to 100% B for 27 min at a flow rate of 0.4 ml min^{-1}, followed by an isocratic flow of 100% B for 3 min. The spectrometer settings were: capillary voltage = 4.0 kV, drying gas temperature = 350 °C at 10 l min^{-1}, and the nebulizer pressure was 45 p.s.i. Metabolite peak transitions and retention times are listed in Extended Data Table 2 and were confirmed by running standards for measured glycolytic, pentose phosphate pathway, tricarboxylic acid cycle and co-factor metabolites. 2-phosphoglycerate and 3-phosphoglycerate isomers were quantified in aggregate. Relative metabolite abundance was quantified by integrated peak area for the given multiple reaction monitoring transition, and all metabolite levels were normalized to internal standard extracted ion intensity values for d_3-serine. These parameters were used to quantify all metabolites, with the exception of 1,3-BPG and MGO, which required chemical derivitization to stable intermediates before LC–MS/MS quantification, as previously reported[20,33]. MGO deviated from all other metabolites, as it was separated on a Gemini reverse-phase C18 column (5 mm, 4.6 mm × 50 mm; Phenomenex) together with a pre-column (C18, 3.5 mm, 2 mm × 20 mm) and detected in positive mode analysis, with mobile phase A (H_2O) and B (50:50 (v/v) H_2O:CH_3CN) supplemented with 0.1% trifluoroacetic acid. The gradient started with 0% B for 2 min and increased linearly to 100% B over 10 min with a flow rate of 0.4 ml min^{-1}, followed by an isocratic gradient of 100% B for 5 min at 0.4 ml min^{-1}. The QQQ settings were the same as above.

Flag-tagged protein expression and western blotting. Full-length, human PGK1 (NM_000291, Origene) transiently expressed from a pCMV6 entry vector with a C-terminal Myc-DDK tag; full-length, human KEAP1 (28023, Addgene) was transiently expressed from a pcDNA/FRT/TO plasmid with a C-terminal 3×Flag tag. All references to Flag–PGK1 or Flag–KEAP1 represent the proteins in the aforementioned vectors, respectively. Transient protein expression was performed in confluent 10 cm plates of HEK293T cells by transfection of 1 μg plasmid with Lipofectamine 2000 (Invitrogen) according to manufacturer's protocol. For in situ compound or metabolite treatment experiments, compounds were added approximately 24 h after transfection, and incubated for the indicated duration. For Flag–KEAP1 western blotting and immunoprecipitation experiments, cells were collected by scraping, pelleted by centrifugation, washed twice with PBS and lysed in 8 M urea, 50 mM NH_4HCO_3, phosphatase inhibitor cocktail (Sigma Aldrich), and EDTA-free complete protease inhibitor (Roche), pH 8.0, at 4 °C. Lysate was sonicated (Fisher Scientific, FB-505), insoluble debris cleared by centrifugation, and the supernatant was diluted into 4× Laemmli buffer containing 50 mM dithiothreitol (DTT) as a reducing agent. Samples were prepared for SDS–PAGE by heating to 95 °C for 5 min, cooled to room temperature, resolved on NuPAGE Novex 4–12% Bis-Tris Protein Gels (Invitrogen), and transferred onto nitrocellulose membranes by standard western blotting methods. Membranes were blocked in 2% BSA in TBS containing 0.1% tween-20 (TBST) and probed with primary and secondary antibodies. Primary antibodies used in this study include: anti-Flag-M2 (1:1,000, F1804, Sigma Aldrich), anti-KEAP1 (1:500, SC-15246, Santa Cruz), anti-HSPA1A (1:1,000; 4872, Cell Signaling), anti-ACTB (1:1,000, 4790, Cell Signaling), anti-GAPDH (1:1,000; 2118S, Cell Signaling) and TUBG (1:1,000; 5886, Cell Signaling). Rabbit polyclonal anti-pgK antibody was generated using pgK-modified KLH and affinity purification as described[4] at a 1:400 dilution of a 0.33 mg ml^{-1} stock in 10 mM sodium HEPES (pH 7.5), 150 mM NaCl, 30% glycerol and 0.02% sodium azide. Secondary donkey anti-rabbit, donkey anti-goat, and donkey anti-mouse (Licor), were used at 1:10,000 dilution in 2% BSA-containing TBST and incubated for 1 h before washing and imaging on a Licor infrared scanner. Densitometry measurements were performed with ImageJ software.

Time- and dose-dependent CBR-470-1 treatment studies were performed in HEK293T cells 24 h after transient transfection of Flag–KEAP1, or in IMR32 cells for endogenous KEAP1. Fresh RPMI media with 10% FBS, 2 mM L-glutamine, 1% penicillin/streptomycin and the indicated concentration of CBR470-1 (20 μM for time-dependent experiments) or equivalent DMSO was added to cells in 10 cm dishes. After the indicated incubation time, cells were lysed in lysis buffer (50 mM Tris, 150 mM NaCl, 1% Triton-X 100, phosphatase inhibitor cocktail (Sigma

Aldrich), and EDTA-free complete protease inhibitor (Roche), pH 7.4) and processed for western blot as indicated above.

Target identification studies with CBR-470-PAP. Confluent IMR32 cells in 10 cm dishes were exposed to 5 μM CBR-470-PAP with either DMSO or a 50-fold molar excess of CBR-470-1 (250 μM) for 1 h at 37 °C. Samples were then UV-crosslinked using a Stratalinker 2400 instrument for 10 min. RIPA-extracted lysates were then fractionated with ammonium sulfate with 20% increments. These fractions were then separated via SDS–PAGE and relevant probe labelling was determined by anti-biotin (1:500, ab1227, Abcam) western blotting as above. A parallel gel was silver stained using the Pierce silver stain kit. Relevant gel slices from the 80% fraction were excised and PGK1 identity was determined by LC–MS/MS by the Scripps Center for Metabolomics and Mass Spectrometry. Follow-up shRNA knockdown studies confirmed PGK1 as the target within this fraction.

Dye-based thermal denaturation assay. Thermal denaturation experiments were performed using a Protein Thermal Shift Dye Kit (ThermoFisher, 4461146). Reactions contained 2 μM recombinant PGK1 with the indicated dose of aqueously delivered CBR-470-1 with 1× supplied thermal shift dye and reaction buffer in 20 μl reaction volumes. Fluorescence values were recorded using a Viia7 Real-Time PCR instrument according to supplied instructions.

Recombinant PGK1 assay. PGK1 enzymatic activity in the forward direction was measured with a coupled enzymatic assay[34]. Three PGK1 conditions were prepared by dissolving recombinant PGK1 in potassium phosphate buffer (10 mM KH_2PO_4, 10 mM $MgSO_4$, pH 7.0), and transferring the aliquots of PGK1 solution to the microtubes being treated with same amount of DMSO and indicated concentrations of CBR-470-1. Final concentration of PGK1 is 20 ng ml^{-1} and DMSO is 1% for each sample. Two blank conditions, 0 μM and 100 μM of CBR-470-1 without PGK1, were also prepared for the control measurements. All PGK1 samples and blank samples were pre-incubated for 20 min and then transferred to the UV-transparent 96 well plate (Corning). The assay solution (10 mM KH_2PO_4, 2 mM G3P, 0.6 mM NAD^+, 200 mM glycine, 0.4 mM ADP, pH 7.0) was activated by adding GAPDH with 10 μg ml^{-1} final concentration, and then the assay solution was added to the wells containing PGK1 samples and blank samples. The change in absorbance at 340 nm at room temperature was measured every 20 s for 45 min, by Tecan Infinite M200 plate reader. Each condition was performed with three independent replications.

Isothermal dose response profiling of PGK1. In vitro thermal profiling assay for recombinant proteins was performed by dissolving pure recombinant PGK1 and GAPDH into PBS, and dividing equal amount of mixture into 9 aliquots. Each aliquot was transferred to 0.2 ml PCR microtubes being treated with different amounts of CBR-470-1 added from DMSO stock, and equal amount of DMSO for the control. Each microtube contains 50 μl of mixture with final concentration of 45 μg ml^{-1} for each protein and DMSO concentration 1% with following final concentrations of CBR-470-1; 0 μM, 0.1 μM, 0.3 μM, 1 μM, 3 μM, 10 μM, 33 μM, 100 μM and 333 μM. After a 30 min incubation at 25 °C, samples were heated at 57 °C for 3 min followed by cooling at 25 °C for 3 min using Thermal Cycler. The heated samples were centrifuged at 17,000*g* for 20 min at 4 °C, and the supernatants were transferred to new Eppendorf tubes. Control experiments were performed with heating at 25 °C for 3 min, instead of 57 °C. Samples were analysed by SDS–PAGE and western blot.

Metabolite treatments and HMM-KEAP1 screening. For in vitro screening of glycolytic metabolites, HEK293T cells expressing Flag–KEAP1 were lysed by snap-freeze-thaw cycles (3×) in PBS, pH 7.4, containing EDTA-free complete protease inhibitor (Roche). Lysates were cleared by centrifugation and the supernatants normalized for concentration by Bradford reagent (2 mg ml^{-1}). Concentrated stocks of each metabolite were made in PBS, which were added to the lysate samples for the final indicated concentrations and incubated at 37 °C for 2.5 h with shaking. After incubation, samples were denatured with 6 M urea and processed for SDS–PAGE and western blotting. Methylglyoxal (40% (v/v) with H_2O), GAP, dihydroxyacetone phosphate, and 2,3-BPG were all obtained from Sigma Aldrich and used as PBS stocks. In situ metabolite treatments were performed in HEK293T cells 24 h after transfection of Flag–KEAP1, treated with MGO (1 or 5 mM) in H_2O (Sigma) or equivalent vehicle alone for 8 h. Cells were collected by scraping, washed in PBS and centrifuged, and lysed in urea lysis buffer and analysis by SDS–PAGE and western blot. Dose–response experiments were performed with high purity MGO was prepared by acidic hydrolysis of MG-1,1-dimethylacetal (Sigma Aldrich) followed by fractional distillation under reduced pressure and colorimetric calibration of the distillates, as previously reported[33]. For in vitro MGO dose–response dimerization of KEAP1, HEK293T cells expressing Flag–KEAP1 were lysed in PBS as indicated above, then serial dilutions of high purity MGO in 50 mM sodium phosphate, pH 7.4, were added to the equal volume of lysate aliquots with final protein concentration of 1 mg ml^{-1}. Each mixture was incubated at 37 °C for 8 h with rotating, HMM-KEAP1 formation was analysed by SDS–PAGE and western blot.

For studies with recombinant KEAP1, Flag–KEAP1 was expressed in HEK293T cells from transient transfection of the Flag–KEAP1 plasmid (Addgene plasmid 28023). Flag–KEAP1 protein was immunopurified after overnight incubation at 4 °C with anti-Flag M2 magnetic beads (Sigma) in RIPA buffer in the presence of protease inhibitors, eluted with 3× Flag peptide (150 ng ml^{-1}) in PBS, and desalted completely into PBS. 500 ng of purified Flag–KEAP1 protein was then subjected to reducing conditions with the addition of either TCEP (0.1 mM) or DTT (1 mM) for 10 min at 37 °C. MGO was then added to a final concentration of 5 mM and incubated for 2 h at 37 °C. Reactions were quenched by the addition of 50 μl of 4× sample buffer and subsequent boiling for 10 min. 12 μl of this reaction was then separated by SDS–PAGE and the presence of HMM-KEAP1 evaluated by anti-Flag western blotting as described or by silver staining using the Pierce Silver Stain Kit (ThermoFisher Scientific).

Site-directed mutagenesis of KEAP1. KEAP1 mutants were generated with PCR primers in Extended Data Table 1 according to the Phusion site-directed mutagenesis kit protocol (F-541, Thermo Scientific) and the QuikChange site-directed mutagenesis kit protocol (200523, Agilent). Mutant KEAP1 plasmids were verified by sequencing (CMV (forward), wild-type primers in the middle of the *KEAP1* sequence (forward) and BGH (reverse)), and were transiently expressed in HEK293T cells in the same manner as wild-type KEAP1. Screening of CBR-470-1-induced HMM-KEAP1 formation with mutant constructs was performed just as with wild-type KEAP1, after 8 h CBR-470-1 treatment (20 μM). After treatment, cells were collected and prepared for SDS–PAGE and western blotting as indicated above.

SILAC cell culture methods and proteomic sample preparation. SILAC labelling was performed by growing cells for at least five passages in lysine- and arginine-free SILAC medium (RPMI, Invitrogen) supplemented with 10% dialysed fetal calf serum, 2 mM L-glutamine and 1% penicillin/streptomycin. 'Light' and 'heavy' media were supplemented with natural lysine and arginine (0.1 mg ml^{-1}), and ^{13}C-, ^{15}N-labelled lysine and arginine (0.1 mg ml^{-1}), respectively.

General protein digestion for LC–MS/MS analysis was performed by dissolving protein (for example, whole lysate or enriched proteins) in digestion buffer (8 M urea, 50 mM NH_4HCO_3, pH 8.0), followed by disulfide reduction with DTT (10 mM, 40 min, 50 °C), alkylation (iodoacetamide, 15 mM, 30 min, room temperature, protected from light) and quenching (DTT, 5mM, 10 min, room temperature). The proteome solution was diluted fourfold with ammonium bicarbonate solution (50 mM, pH 8.0), $CaCl_2$ added (1 mM) and digested with sequencing grade trypsin (~1:100 enzyme/protein ratio; Promega) at 37 °C while rotating overnight. Peptide digestion reactions were stopped by acidification to pH 2–3 with 1% formic acid, and peptides were then desalted on ZipTip C18 tips (100 μl, Millipore), dried under vacuum, resuspended with LC–MS grade water (Sigma Aldrich), and then lyophilized. Lyophilized peptides were dissolved in LC–MS/MS buffer A (H_2O with 0.1% formic acid, LC–MS grade, Sigma Aldrich) for proteomic analysis.

Proteomic LC–MS/MS and data analysis. LC–MS/MS experiments were performed with an Easy-nLC 1000 ultra-high pressure LC system (ThermoFisher) using a PepMap RSLC C18 column heated to 40 °C (column: 75 μm × 15 cm; 3 μm, 100 Å) coupled to a Q Exactive HF orbitrap and Easy-Spray nanosource (ThermoFisher). Digested peptides (500 ng) in MS/MS buffer A were injected onto the column and separated using the following gradient of buffer B (0.1% formic acid acetonitrile) at 300 nl min^{-1}: 0–2% buffer B over 10 min, 2–40% buffer B over 120 min, 40–70% buffer B over 10 min, and 70–100% buffer B over 5 min. MS/MS spectra were collected from 0 to 150 min using a data-dependent, top-20 ion setting with the following settings: full MS scans were acquired at a resolution of 120,000, scan range of 400–1,600 *m*/*z*, maximum IT of 50 ms, AGC target of 1×10^6, and data collection in profile mode. MS/MS scans were performed by HCD fragmentation with a resolution of 15,000, AGC target of 1×10^5, maximum IT of 30 ms, NCE of 26, and data type in centroid mode. Isolation window for precursor ions was set to 1.5 *m*/*z* with an underfill ratio of 0.5%. Peptides with charge state >5, 1 and undefined were excluded and dynamic exclusion was set to eight seconds. Furthermore, S-lens RF level was set to 60 with a spray voltage value of 2.60 kV and ionization chamber temperature of 300 °C.

MS/MS files were generated and searched using the ProLuCID algorithm in the Integrated Proteomics Pipeline (IP2) software platform. Human proteome data were searched using a concatenated target/decoy UniProt database (UniProt_Human_reviewed_04-10-2017.fasta). Basic searches were performed with the following search parameters: HCD fragmentation method; monoisotopic precursor ions; high resolution mode (3 isotopic peaks); precursor mass range 600–6,000 and initial fragment tolerance at 600 p.p.m.; enzyme cleavage specificity at C-terminal lysine and arginine residues with three missed cleavage sites permitted; static modification of +57.02146 on cysteine (carboxyamidomethylation); two total differential modification sites per peptide, including oxidized methionine (+15.9949); primary scoring type by XCorr and secondary by Zscore; minimum peptide length of six residues with a candidate peptide threshold of 500. A minimum of one peptide per protein and half-tryptic peptide specificity were required. Starting statistics were performed with a Δmass cutoff = 15 p.p.m. with modstat, and trypstat settings. False discovery rates of peptide (sfp) were set to 1%, peptide

modification requirement (-m) was set to 1, and spectra display mode (-t) was set to 1. SILAC searchers were performed as above with light and heavy database searches of MS1 and MS2 files by including static modification of +8.014168 for lysine and +10.0083 for arginine in a parallel heavy search. SILAC quantification was performed using the QuantCompare algorithm, with a mass tolerance of 10 p.p.m. or less in cases in which co-eluting peptide interferes. In general, all quantified peptides have a mass error within 3 p.p.m.

Quantitative proteomic detection of potential KEAP1 modification sites. Quantitative surface mapping with SILAC quantitative proteomics was performed with heavy- and light-labelled HEK293T cells expressing Flag–KEAP1. Cells were incubated with DMSO alone (light cells) or CBR-470-1 (20 μM, heavy cells) for 8 h. After incubation cells were scraped, washed with PBS (3×) and combined before lysis in urea lysis buffer (8 M urea, 50 mM NH_4HCO_3, nicotinamide (1 mM), phosphatase inhibitor cocktail (Sigma Aldrich), and EDTA-free complete protease inhibitor (Roche), pH 8.0) by sonication at 4 °C. After sonication insoluble debris was cleared by centrifugation (17,000*g*, 10 min), diluted with Milli-Q water to give 1 M urea, and lysate was incubated with anti-Flag M2 resin (100 μl slurry, A2220, Sigma Aldrich) at 4 °C overnight while rotating. For SILAC label-swap experiments, light HEK293T cells were incubated with CBR-470-1 and heavy cells were incubated with DMSO and processed as above. Flag resin was washed with PBS (7 × 1 ml), Flag–KEAP1 protein eluted with glycine-HCl buffer (0.1 M glycine, pH 3.5, 2 × 500 μl), followed by 8 M urea (2 × 100 μl). The combined eluent was brought up to 8 M urea total concentration and processed for trypsin digestion and LC–MS/MS analysis as indicated above.

The SILAC maps were generated by comparing SILAC ratios for each peptide, relative to the median value for all KEAP1 peptides. SILAC ratios were converted to Log_2 values and plotted to visualize peptides that are significantly perturbed, for example by modification, relative to the rest of the protein. A minimum of three SILAC ratios for each peptide was required for inclusion in KEAP1 surface maps, which allowed for approximately 85–90% coverage of the KEAP1 protein. Missing sequences were caused by the lack or close spacing of tryptic sites, resulting in inadequate peptides for MS/MS detection.

In vitro MGO peptide reactions. 'CR' peptide was synthesized using standard solid phase peptide synthesis with FMOC-protected amino acids on MBHA rink amide resin. Peptides were cleaved in a solution of 94% trifluoroacetic acid, 2.5% triisopropyl silane, 2.5% H_2O, 1% β-mercaptoethanol and precipitated with ether. Peptide identity was confirmed using an Agilent 1100 series LC–MS. Peptides were purified via reverse phase HPLC on an Agilent Zorbax SB-C18 250 mm column and dried via lyophilization. For MGO reactions CR peptide (1 mM) was incubated with 12.5 mM MGO (diluted from 40% solution in water; Sigma Aldrich) or equivalent amount of water (mock) in 1× PBS pH 7.4 at 37 °C overnight. Reactions were diluted 1:25 in 95%/5% H_2O/acetonitrile and 0.1% trifluoroacetic acid and analysed by LC–MS.

For NMR experiments, approximately 1.5 mg of the CR or CR-MGO crosslinked peptide was purified by reverse phase HPLC, lyophilized and dissolved in 700 μl d_6-DMSO. The peptides were dried via lyophilization. All NMR experiments were performed on a Bruker Avance II+ 500 MHz 11.7 Tesla NMR. Data were processed and plotted in Bruker Topspin 3.5. CR peptide NMR experiments were run with a spectral width of 8.5 for 2D experiments (in both dimensions) and 15 for 1D proton NMR with a pulse width of 13.5 μs and an interscan delay of 3 s. For the proton NMR, 256 scans were taken. For the COSY-DQF experiment, 128 and 2,048 complex points were acquired in the F_1 and F_2 dimensions respectively, with 8 scans per point. For the TOCSY experiment, a mixing time of 60 μs was used, and 256 and 1,024 complex points were acquired with 8 scans per point. All CR-MGO peptide NMR experiments were run with a spectral width of 13 (in both dimensions) with a pulse width of 11.5 μs and an interscan delay of 2.2 s. For the proton NMR, 256 scans were taken. For the COSY-DQF experiment, 128 and 2,048 complex points were acquired in the F_1 and F_2 dimensions, respectively, with 8 scans per point. For the TOCSY experiment, a mixing time of 80 μs was used, and 256 and 1,024 complex points were acquired with 8 scans per point.

In-gel digestion of KEAP1. Targeted proteomic analyses of KEAP1 protein were performed by running anti-Flag enriched HMM-KEAP1 and low-molecular mass (LMM)-KEAP1 (from both CBR-470-1 or MGO treatments as above) on SDS–PAGE gels, and isolated gel pieces were digested in-gel with sequencing grade trypsin (Promega), as previously reported[35]. Tryptic peptides from in-gel tryptic digestions were dissolved in 100 mM Tris-HCl, pH 8.0, with 2 mM of $CaCl_2$, and further digested with mass spectrometry-grade chymotrypsin (Thermo Scientific) according to manufacturer's protocol. Chymotryptic digestion reactions were stopped by acidification, and desalted on Ziptip C18 tips.

Targeted proteomic analysis of crosslinked KEAP1 peptides. Double-digested KEAP1 peptides from isolated HMM-KEAP1 and monomeric KEAP1 were analysed by LC–MS/MS on an Easy-nLC 1000 ultra-high pressure LC system coupled to a Q Exactive HF orbitrap and Easy-Spray nanosource as indicated above. Candidate peptides were initially searched by manual inspection of chromatograms and MS1 spectra for *m/z* values of peptide candidates from predicted digestion sites, crosslink sites and differential presence in HMM- and monomeric KEAP1 from both CBR-470-1-and MGO-treated samples. Extracted MS1 ions of the candidates were present in HMM-KEAP1 digests but not in LMM-KEAP1 digests. MS/MS spectra and PRM experiments were collected on the same instrument using the following settings: global and general settings included lock masses of off, chromatography peak width of 15 s, polarity of positive, in-source CID of 0.0 eV, inclusion list set to 'on,' and an *m/z* value of the target parent ion with its charge state in the inclusion list. MS2 scans were performed by HCD fragmentation with microscans of 1, resolution of 120,000, AGC target of 5×10^5, maximum IT of 200 ms, loop count of 1, MSX count of 1, isolation window of 2.0 *m/z*, isolation offset of 0.0 *m/z*, NCE of 16, and spectrum data type in profile mode. Furthermore, S-lens RF level was set to 60 with a spray voltage value of 2.20 kV and ionization chamber temperature of 275 °C. Targeted PRM experiments were performed on CBR-470-1-, MGO-induced HMM-KEAP1 and monomeric KEAP1 samples.

UVB skin damage model. Thirty-two 5-week-old BALB/c male mice were randomized into 4 groups of 8 animals such that each group had similar body weight means. Mice were prepared for removal of hair from their entire back two days before UVB exposure (day 3) by using an electric shaver and depilatory cream. On day 5, mice received exposure to UVB (200 mJ cm^{-2}) produced by a broad band UVB lamp (Dermapal UVB Rev 2) powered by a Kernel UV Phototherapy system. UVB exposure was confined to a rectangular area of ~8 cm^2 by a lead shielding mask. UVB doses were confirmed by dosimeter measurements (Daavlin X96). Sham animals were shaved but received no UVB treatment. Mice were dosed from day 0 to study end at day 10 via oral gavage twice daily (CBR-470-2, 50 mg kg^{-1} twice daily, orally; BARD, 3 mg kg^{-1} orally; vehicle, 0.5% methyl cellulose/0.5% Tween80). Mice were monitored daily for body weight changes and erythema scoring from days 5 to 10. Mice were euthanized at day 10 and specimens collected for histological analysis from the wounded area. These studies were performed at Biomodels, LLC. Blinded erythema scores were recorded by a blinded, trained investigator according to established in house scale. In short, a scale of 0 to 4 was generated with a score of 0 referring to normal skin and a score of 4 indicating severe ulceration.

Percentage wounded area measurements. Photographs of animals on day 10 of the study were taken such that the distances from camera, aperture and exposure settings were identical. Images were then cropped such that only the shaved, wounded area encompassed the imaging field. These images were then processed with a custom ImageJ macro which first performed a three colour image deconvolution to separate the red content of the image[36]. The thresholding function within ImageJ software was then used to separate clear sites of wounding from red background present in normal skin. Red content corresponding to wounds was then quantified as a fraction of the whole imaging field and reported as the percentage wounded area.

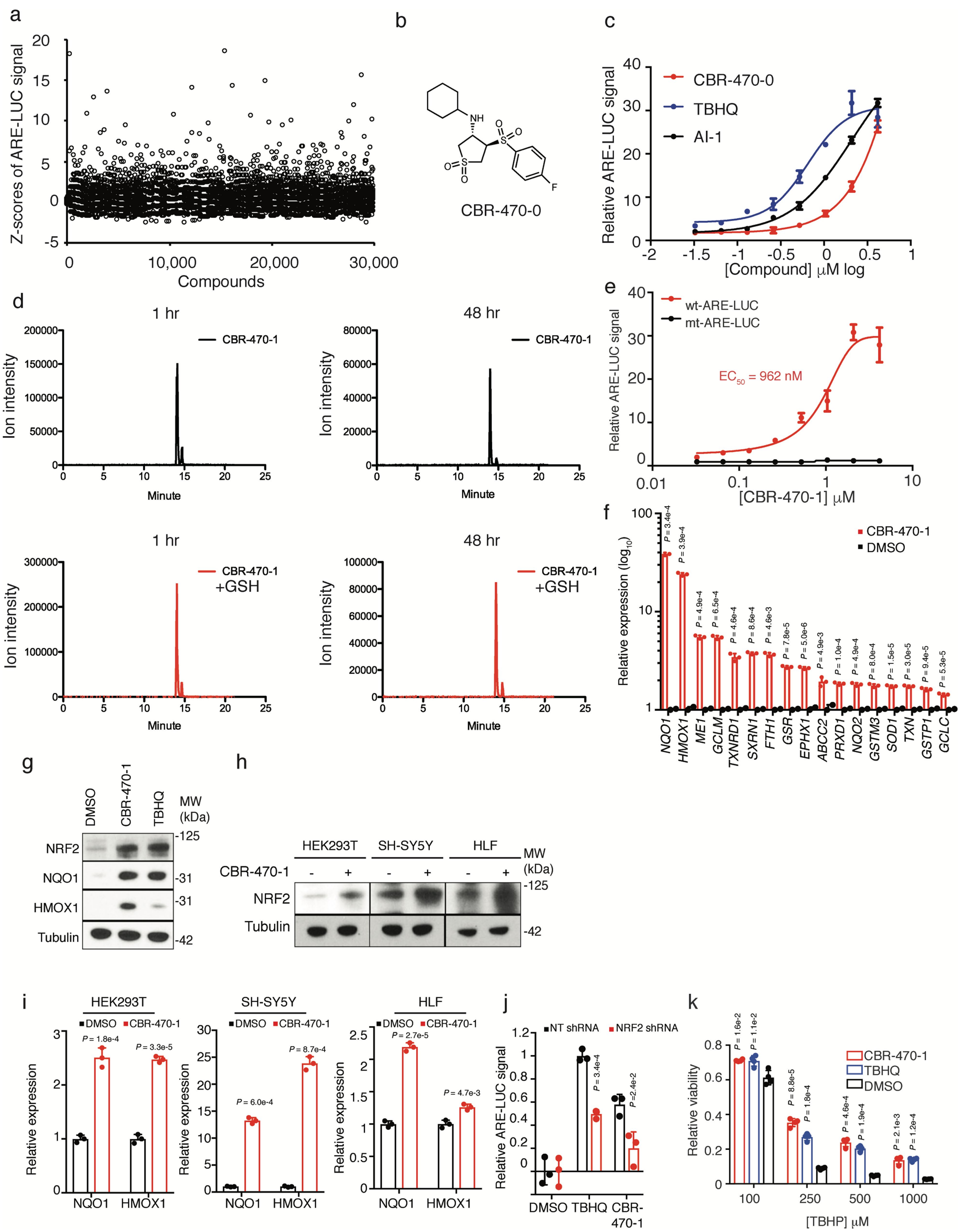
Epidermal thickness measurements. H&E-stained skin sections corresponding to the wounded area were generated by Histotox Labs and accessed via pathxl software. Twenty-four individual measurements of epidermal thickness from 8 sections spanning a 400-μm step distance were recorded per animal by a non-blinded, trained investigator. These measurements were then averaged to generate a mean epidermal thickness measurement per animal.

Reporting summary. Information on research design is available the Nature Research Reporting Summary.

Data availability

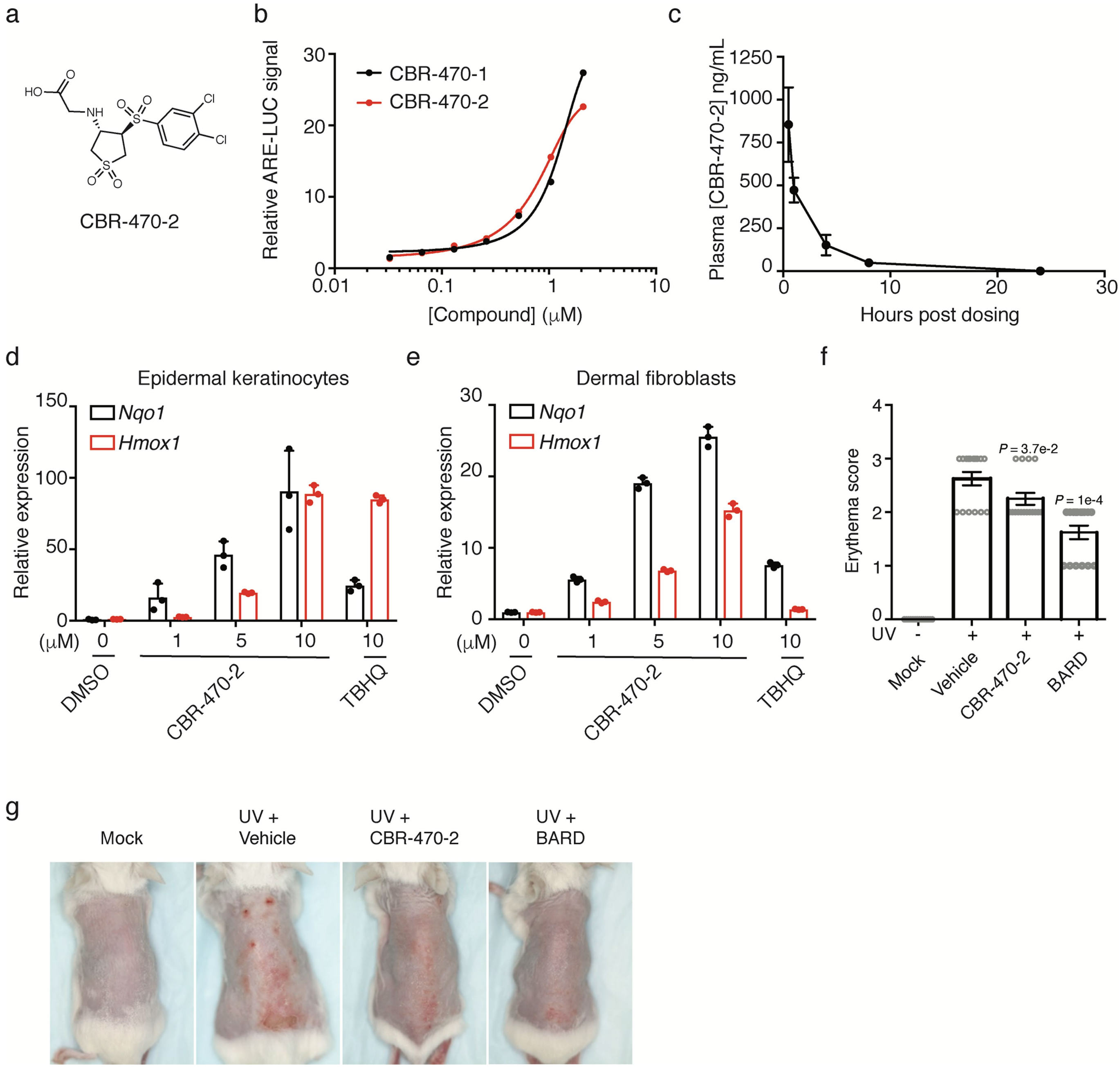
RNA-seq primary data are deposited in the Gene Expression Omnibus (GEO) under accession number GSE116642. Source data for all mouse experiments have been provided. Full scans for western blots and gels are available in the Supplementary Information. All other data are available on reasonable request.

32. Hur, W. & Gray, N. S. Small molecule modulators of antioxidant response pathway. *Curr. Opin. Chem. Biol.* **15**, 162–173 (2011).
33. Rabbani, N. & Thornalley, P. J. Measurement of methylglyoxal by stable isotopic dilution analysis LC–MS/MS with corroborative prediction in physiological samples. *Nat. Protoc.* **9**, 1969–1979 (2014).
34. Bücher, T. Phosphoglycerate kinase from Brewer's yeast. *Methods Enzymol.* **1**, 415–422 (1955).
35. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **1**, 2856–2860 (2007).
36. Ruifrok, A. C. & Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **23**, 291–299 (2001).

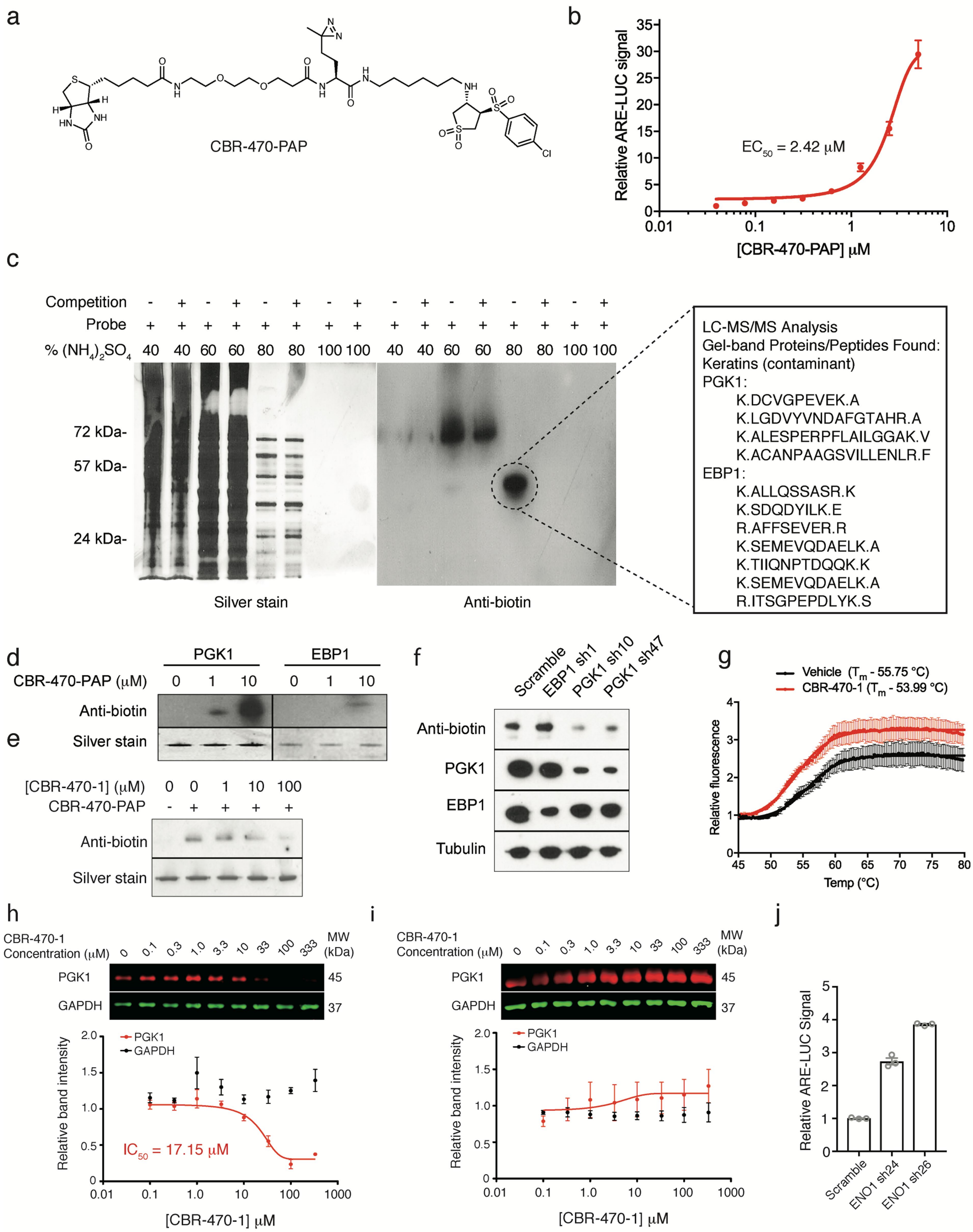


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | A high-throughput screen identifies a non-covalent NRF2 activator chemical series that activate a robust NRF2 transcriptional program in multiple cell types. **a**, Plate-based *Z*-scores of ARE-LUC luminance measurements of all test compounds from a 30,000 compound screen in IMR32 cells. **b**, Structure of screening hit CBR-470-0. **c**, Relative ARE-LUC luminance measurements from IMR32 cells treated for 24 h with a concentration response of CBR-470-0 and reported NRF2 activators TBHQ and AI-1 ($n = 3$ biologically independent samples, mean and s.e.m.). **d**, LC–MS quantification of CBR-470-1 (50 μM) incubated in the presence or absence of GSH (1 mM) in PBS for 1 h (left) and 48 h (right). Relative ion intensities within each time point were compared with representative chromatograms shown ($n = 2$). **e**, Relative ARE-LUC luminance values from IMR32 cells transfected with wild-type (wt) or mutant (mt; two core nucleotides necessary for NRF2 binding were changed from GC to AT) ARE-LUC reporter constructs and treated with the indicated doses of CBR-470-1 for 24 h ($n = 3$, mean and s.e.m.). **f**, Relative abundance of NRF2-dependent transcripts as determined by qRT–PCR from IMR32 cells treated for 24 h with 5 μM CBR-470-1 ($n = 3$). **g**, Western blot analyses of total NRF2 protein content or NRF2-controlled genes (*NQO1* and *HMOX1*) from IMR32 cells treated for 24 h with 5 μM CBR-470-1 ($n = 5$). **h**, Western blot analyses of total NRF2 protein content from the indicated cell types treated for 4 h with 5 μM CBR-470-1 ($n = 3$). **i**, Relative expression levels of NQO1 and HMOX1 from the indicated cell types treated for 24 h with 5 μM CBR-470-1 ($n = 3$, mean and s.d.). **j**, Relative ARE-LUC luminescence values from HEK293T cells transfected with the indicated shRNA constructs and pTI-ARE-LUC and then treated with TBHQ (10 μM) or CBR-470-1 (5 μM) for 24 h ($n = 3$). **k**, Relative viability measurements of SH-SY5Y cells treated with either CBR-470-1 (5 μM) or TBHQ (10 μM) for 48 h and then challenged with the indicated doses of TBHP for 8 h ($n = 4$). Data are mean and s.d. or s.e.m. of biologically independent samples. $*P < 0.05$, $**P < 0.005$, $***P < 0.001$, univariate two-sided *t*-test.

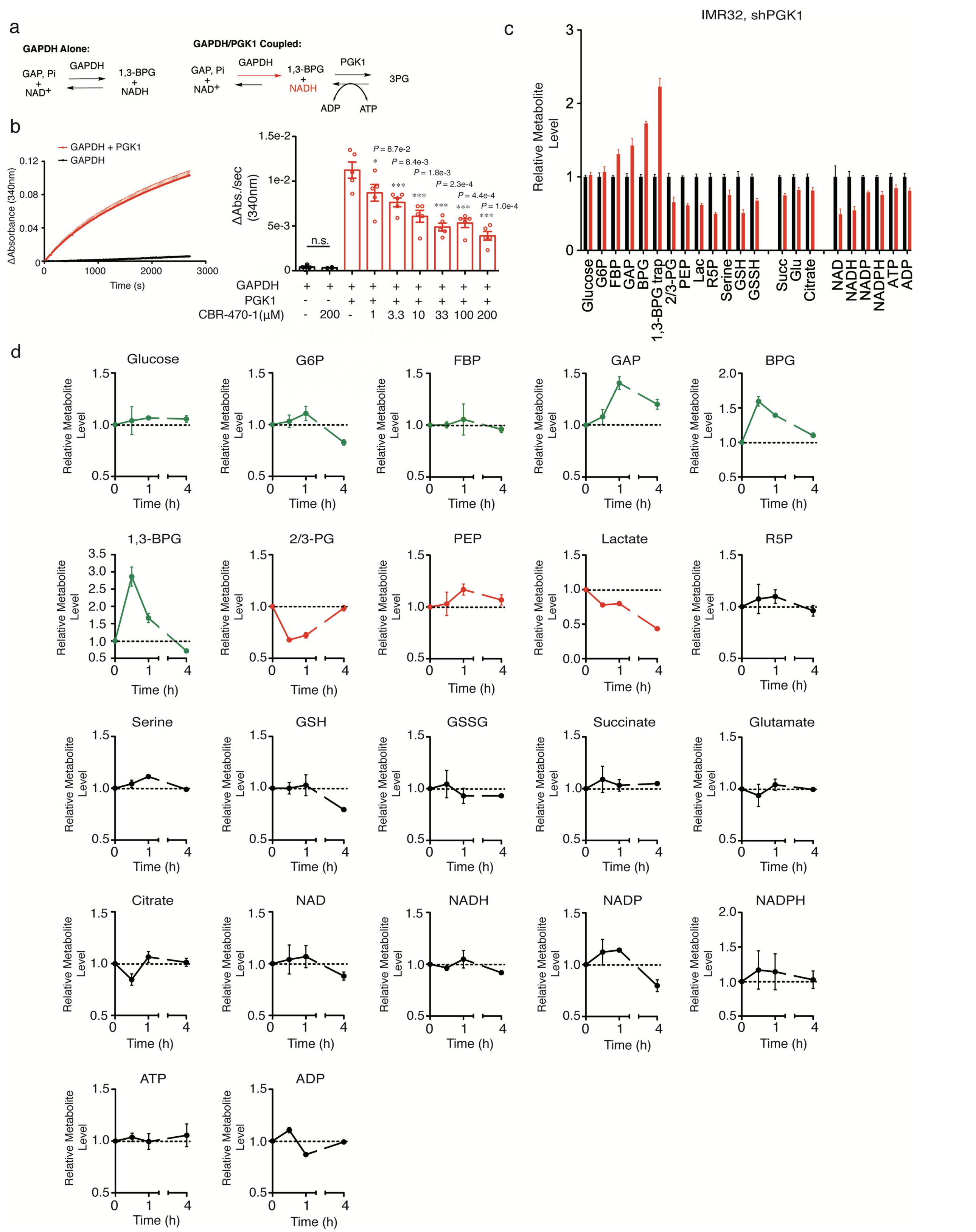


Extended Data Fig. 2 | CBR-470-2 pharmacokinetics and in vivo activity. a, Structure of CBR-470-2. **b**, Relative ARE-LUC luminance values from IMR32 cells transfected with pTI-ARE-LUC and treated with the indicated doses of CBR-470-1 and CBR-470-2 for 24 h ($n = 3$ biologically independent samples). **c**, Plasma concentrations of CBR-470-2 from mice treated with a single 20 mg kg^{-1} dose of compound ($n = 3$ animals, mean and s.e.m.). **d**, **e**, Relative transcript levels of *Nqo1* and *Hmox1* from mouse epidermal keratinocytes (**d**) and mouse dermal fibroblasts (**e**) treated for 24 h with the indicated doses of compound ($n = 3$ biologically independent samples, mean and s.d.). **f**, Blinded erythema scores from mice treated with vehicle, CBR-470-2 or bardoxolone after exposure to UV ($n = 8$ animals, mean and s.e.m.). *$P < 0.05$, *** $P < 0.005$, one-way ANOVA with Dunnett's correction. **g**, Representative images of UV-exposed dorsal regions of animals at day 10 of the study.

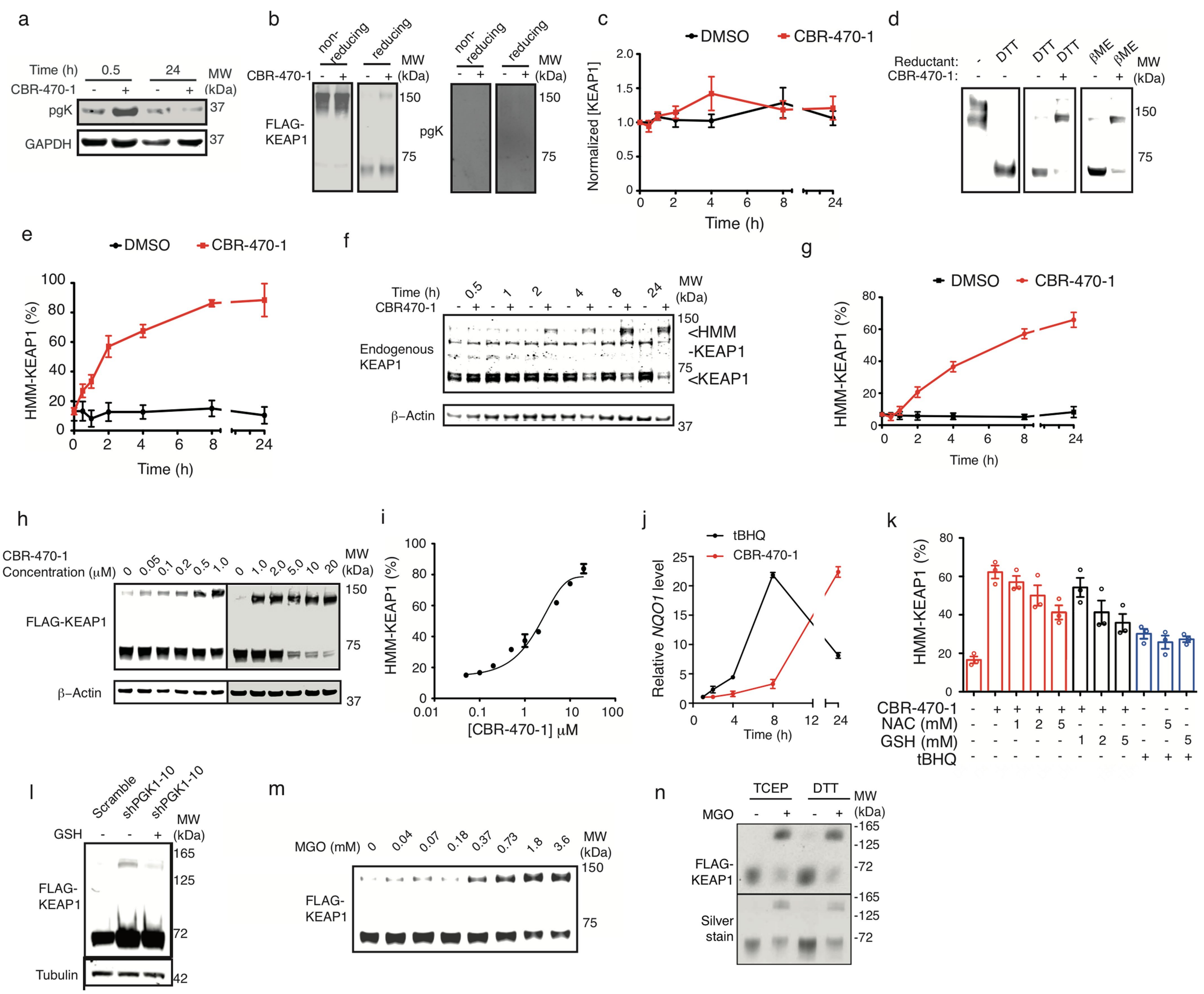


Extended Data Fig. 3 | See next page for caption.

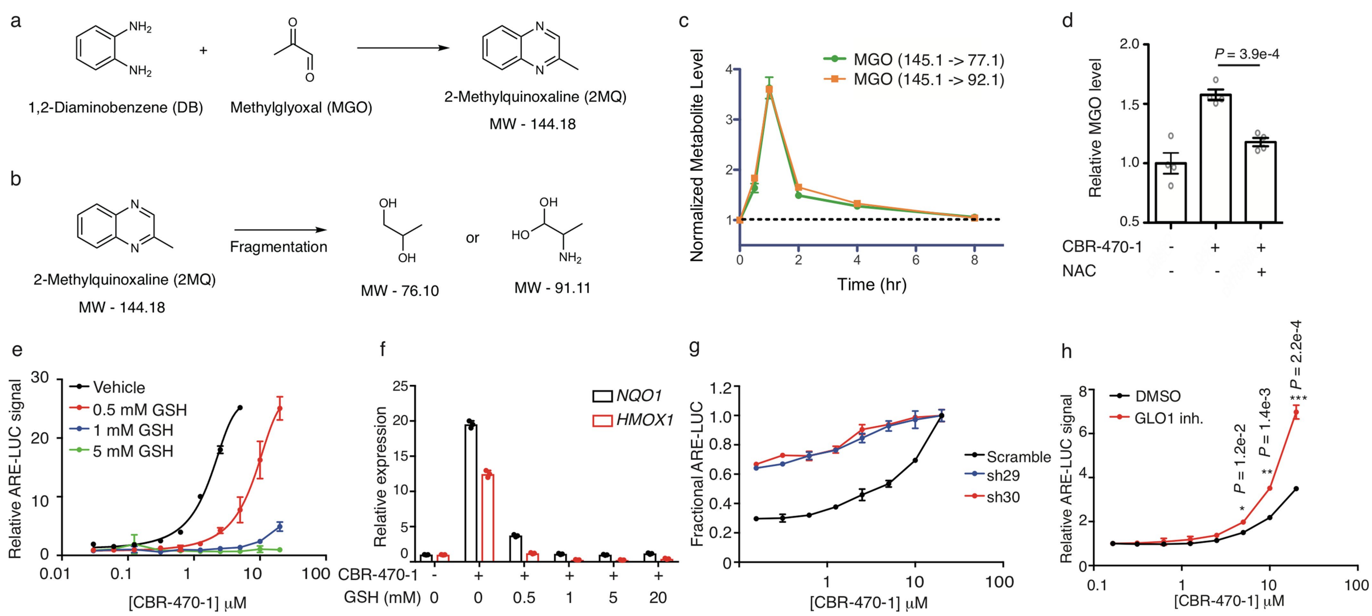
Extended Data Fig. 3 | A photoactivatable affinity probe-based approach identifies PGK1 as the relevant cellular target of CBR-470-1. **a**, Structure of CBR-470-PAP. **b**, Relative ARE-LUC luminance values from IMR32 cells transfected with pTI-ARE-LUC and then treated with the indicated doses of CBR-470-PAP for 24 h ($n = 3$). **c**, Silver staining and anti-biotin western blots of ammonium sulfate fractionated lysates from UV-irradiated IMR32 cells treated with 5 μM for 1 h with or without CBR-470-1 competition (250 μM) ($n = 3$). Shown on the right are initial proteomic target results from gel-band digestion and LC–MS/MS analysis. **d**, Anti-biotin western blots from in vitro crosslinking assays with recombinant PGK1 and EBP1 in the presence of the indicated doses of CBR-470-PAP ($n = 2$). **e**, Anti-biotin western blot analyses from an in vitro crosslinking assay with recombinant PGK1 in the presence of CBR-470-PAP (1 μM) and indicated concentration of soluble CBR-470-1 competitor ($n = 2$). **f**, Anti-biotin western blot analyses of cells treated with 5 μM CBR-470-PAP after transduction with anti-PGK1 and anti-EBP1 shRNA for 48 h. Depletion of PGK1 protein selectively reduces CBR-470-PAP-dependent labelling ($n = 2$). **g**, Dye-based thermal denaturation assay with recombinant PGK1 in the presence CBR-470-1 (20 μM) or vehicle alone ($n = 3$). Calculated melting temperature (T_m) values are listed. **h**, **i**, Dose-dependent thermal stability assay of recombinant PGK1 and GAPDH in the presence of increasing doses of CBR-470-1 near the T_m of both proteins (57 °C) ($n = 5$) (**h**) or room temperature ($n = 3$) (**i**). Western blot of sample supernatants after centrifugation (13,000 r.p.m.) detected total PGK1 and GAPDH protein, which were plotted in Prism (below). **j**, ARE-LUC reporter activity in HEK293T cells with transient shRNA knockdown of *ENO1* ($n = 3$). Data are mean ± s.e.m. of biologically independent samples.



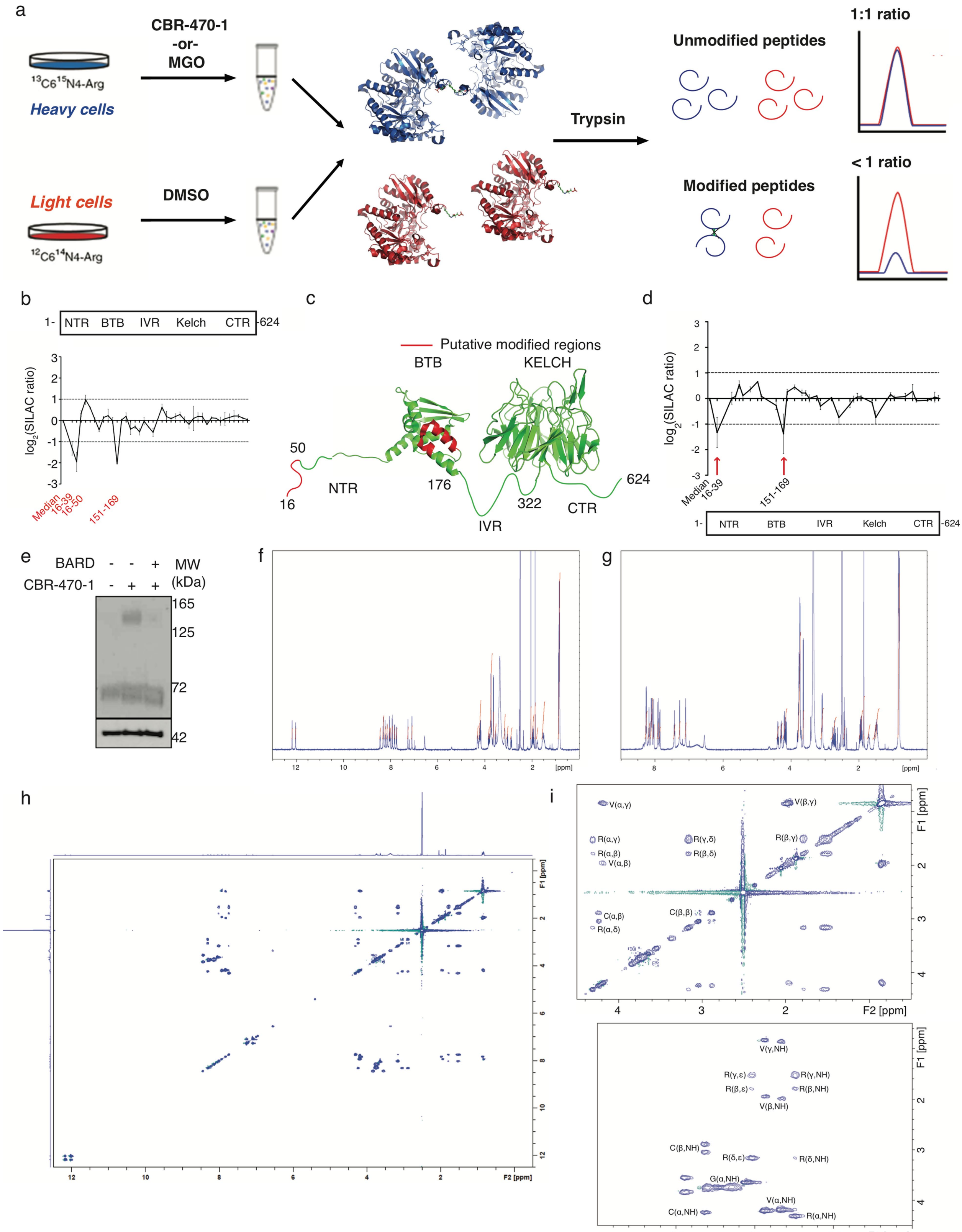
Extended Data Fig. 4 | CBR-470-1 inhibits PGK1 in vitro and in situ. **a**, Schematic of the GAPDH/PGK1 coupled assay. Pre-equilibration of the GAPDH reaction (top left) results in an NAD$^+$/NADH equilibrium, which after addition of PGK1 and ADP pulls the reaction to the right, producing more NADH. Monitoring NADH absorbance after the addition of PGK1 (bottom right) can be used to monitor PGK1 activity in the forward direction (right). Kinetic monitoring of NADH absorbance (340 nm) after established equilibrium with GAPDH shows little change (black curve), but is significantly increased after the addition of PGK1, pulling the equilibrium to the right (red curve). **b**, CBR-470-1 does not affect the GAPDH equilibrium alone, but significantly inhibits PGK1-dependent activity and accumulation of NADH ($n = 5$). **c**, **d**, Relative level of central metabolites in IMR32 cells treated with viral knockdown of PGK1 for 72 h (**c**) ($n = 4$) and with CBR-470-1 relative DMSO alone for the indicative times (**d**) ($n = 3$). Each metabolite is normalized to the control condition at each time point. Statistical analysis was by univariate two-sided t-test (**b**). Data are mean ± s.e.m. of biologically independent samples.



Extended Data Fig. 5 | Modulation of PGK1 induces HMM-KEAP1. **a**, Anti-phosphoglyceryl-lysine (pgK) and anti-GAPDH western blots analysis of CBR-470-1 or DMSO-treated IMR32 cells at early (30 min) and late (24 h) time points ($n=6$). **b**, Anti-Flag (left) and anti-pgK (right) western blot analysis of affinity purified Flag–KEAP1 from HEK293T cells treated with DMSO or CBR-470-1 for 30 min. Duplicate samples were run under non-reducing (left) and reducing (DTT, right) conditions ($n=6$). **c**, Densitometry quantification of total endogenous KEAP1 levels (combined bands at approximately 70 and 140 kDa) in IMR32 cells treated with DMSO or CBR-470-1 for the indicated times ($n=6$). **d**, Western blot detection of Flag–KEAP1 in HEK293T cells comparing non-reducing reagent to DTT (left), and the stability of CBR-470-1-dependent HMM-KEAP1 to the presence of DTT (12.5 mM final concentration, middle) or β-mercaptoethanol (βME; 5% (v/v) final concentration, right) during sample preparation. **e**, Time-dependent CBR-470-1 treatment of HEK293T cells expressing Flag–KEAP1. Time-dependent assays were run with 20 μM CBR-470-1 with western blot analysis at the indicated time-points ($n=8$). **f**, **g**, Western blot detection (**f**) and quantification (**g**) of endogenous KEAP1 and β-actin in IMR32 cells treated with DMSO or CBR-470-1 for the indicated times ($n=6$). Arrows indicate monomeric KEAP1 (70 kDa) and HMM-KEAP1 (140 kDa) bands. **h**, **i**, Western blot (**h**) detection and quantification (**i**) of Flag–KEAP1 in HEK293T cells exposed to increasing doses of CBR-470-1 ($n=3$). **j**, Kinetic qRT–PCR measurement of *NQO1* mRNA levels from IMR32 cells treated with TBHQ (10 μM) or CBR-470-1 (10 μM) for the indicated times ($n=3$). **k**, Quantification of HMM-KEAP1 formation after treatment with CBR-470-1 or the direct KEAP1 alkylator TBHQ, in the presence or absence of GSH or *N*-acetylcysteine (NAC) ($n=3$). All measurements were taken after 8 h of treatment in Flag–KEAP1-expressing HEK293T cells. **l**, Transient shRNA knockdown of *PGK1* induced HMM-KEAP1 formation, which was blocked by co-treatment of cells with GSH ($n=3$). **m**, Anti-Flag western blot analysis of Flag–KEAP1 monomer and the HMM–KEAP1 fraction, with dose-dependent incubation of distilled MGO in lysates from HEK293T cells expressing Flag–KEAP1 ($n=4$). **n**, SDS–PAGE gel (silver stain) and anti-Flag western blot analysis of purified KEAP1 treated with MGO under the indicated reducing conditions for 2 h at 37 °C ($n=3$). Purified protein reactions were quenched in 4× SDS loading buffer containing β-mercaptoethanol and processed for gel analysis as in **d**. Data are mean ± s.e.m. of biologically independent samples.

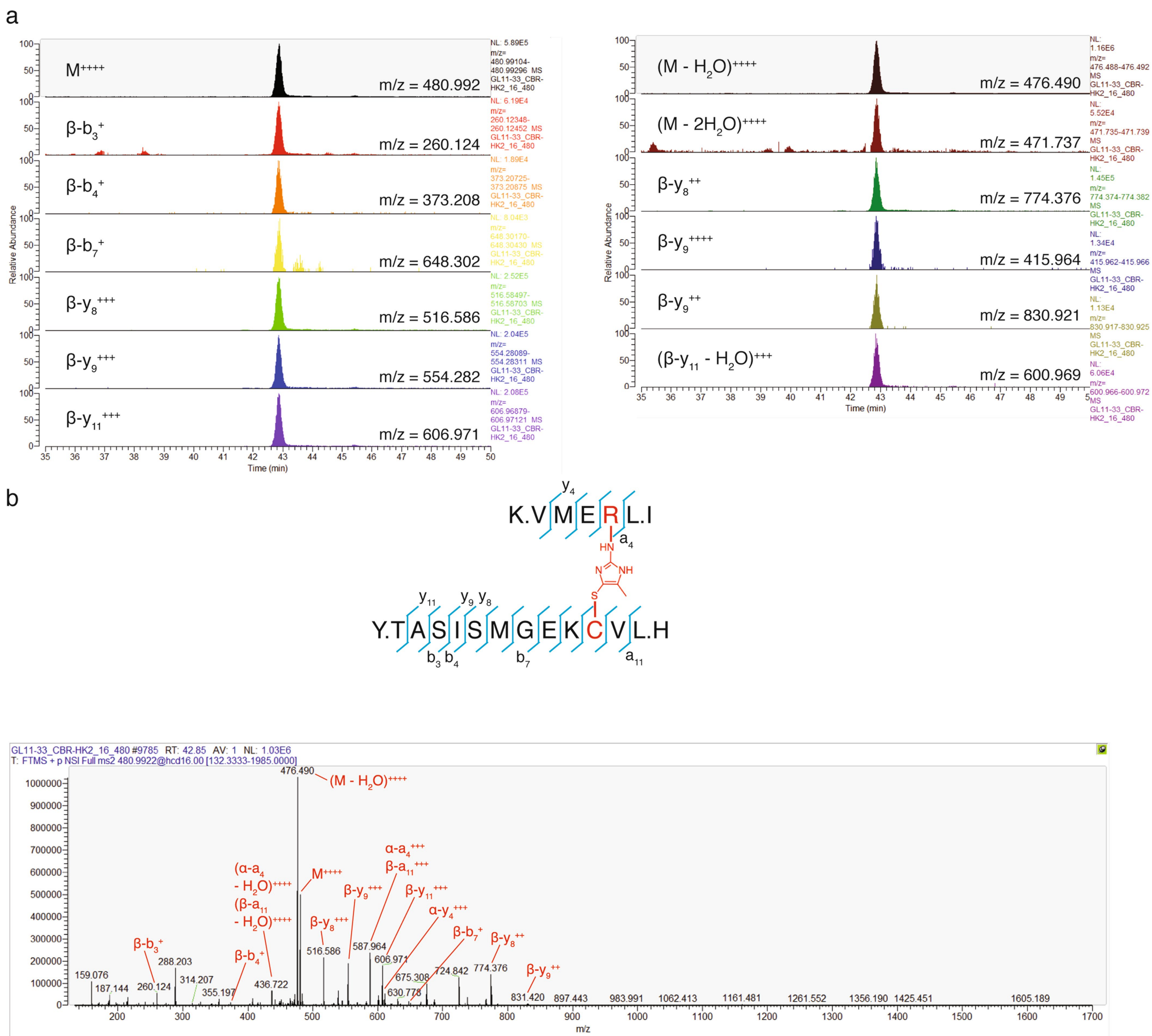


Extended Data Fig. 6 | MGO and glyoxylase activity regulates NRF2 activation. CBR-470-1 causes increased levels of MGO in cells. **a**, Schematic depicting chemical derivatization and trapping of cellular MGO for analysis by targeted metabolomics using two unique fragment ions. **b**, **c**, Daughter ion fragments (**b**) and resulting MS/MS quantification of MGO levels (**c**) in IMR32 cells treated with CBR-470-1, relative to DMSO ($n = 4$). **d**, Quantitative LC–MS/MS measurement of cellular MGO levels in IMR32 cells treated for 2 h with CBR-470-1 or co-treated for 2 h with CBR-470-1 and *N*-acetylcysteine (2 mM) relative to DMSO ($n = 4$). **e**, Relative ARE-LUC luminance values from IMR32 cells transfected with pTI-ARE-LUC and co-treated with the indicated doses of CBR-470-1 and GSH ($n = 3$). **f**, Relative levels of *NQO1* and *HMOX1* transcripts from IMR32 cells co-treated with CBR-470-1 (10 μM) and the indicated concentrations of GSH for 24 h ($n = 3$). **g**, Fractional ARE-LUC values from HEK293T cells transiently co-transfected with pTI-ARE-LUC and the indicated shRNAs and then treated for 24 h with the indicated doses of CBR-470-1 ($n = 3$). **h**, ARE-LUC reporter activity in HEK293T cells treated with CBR-470-1 alone (black) and with a cell-permeable small-molecule GLO1 inhibitor (red) ($n = 3$). Statistical analysis was by univariate two-sided *t*-test (**d**, **h**). Data are mean ± s.e.m. of biologically independent samples.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Schematic of SILAC-based proteomic mapping of KEAP1 modifications in response to CBR-470-1 and NMR characterization of CR-MGO peptide. a, SILAC experiments in which Flag-tagged KEAP1 was treated with vehicle ('light') and CBR-470-1 or MGO ('heavy'). Subsequent mixing of the cell lysates, anti-Flag enrichment, tryptic digestion and LC–MS/MS analysis permitted the detection of unmodified portions of KEAP1, which retained approximate 1:1 SILAC ratios relative to the median ratios for all detected KEAP1 peptides. By contrast, peptides that are modified under one condition will no longer match tryptic MS/MS searches, resulting in skewed SILAC ratios that 'drop out' (bottom). **b**, SILAC ratios for individual tryptic peptides from Flag–KEAP1-enriched DMSO-treated light cells and CBR-470-1-treated heavy cells, relative to the median ratio of all KEAP1 peptides. Highlighted tryptic peptides were significantly reduced by three- to fourfold relative to the KEAP1 median, indicative of structural modification ($n = 8$). **c**, Structural depiction of potentially modified stretches of human KEAP1 (red) using published X-ray crystal structure of the BTB (PDB code 4CXI) and KELCH (PDB code 1U6D) domains. Intervening protein stretches are depicted as unstructured loops in green. **d**, SILAC ratios for individual tryptic peptides from Flag–KEAP1-enriched MGO-treated heavy cell lysates and non-treated light cell lysates, relative to the median ratio of all KEAP1 peptides. Highlighted tryptic peptides were significantly reduced by 2- to 2.5-fold relative to the KEAP1 median, indicative of structural modification ($n = 12$). **e**, Representative western blotting analysis of Flag–KEAP1 dimerization from HEK293T cells pre-treated with BARD followed by CBR-470-1 treatment for 4 h ($n = 3$). **f**, ^{1}H-NMR of CR-MGO peptide (isolated product of MGO incubated with Ac-NH-VVCGGGRGG-C(O)NH_2 peptide). ^{1}H-NMR (500 MHz, d_6-DMSO) δ 12.17 (s, 1H), 12.02 (s, 1H), 8.44 (t, J = 5.6 Hz, 1H), 8.32–8.29 (m, 2H), 8.23 (t, J = 5.6 Hz, 1H), 8.14 (t, J = 5.9 Hz, 1H), 8.05 (t, J = 5.9 Hz, 1H), 8.01 (t, J = 5.9 Hz, 1H), 7.93 (d, J = 8.5 Hz, 1H), 7.74 (d, J = 8.0 Hz, 1H), 7.26 (s, 1H), 7.09 (s, 1H), 4.33–4.28 (m, 1H), 4.25-4.16 (m, 3H), 3.83 (dd, J = 6.9 Hz, J = 16.2 Hz, 1H), 3.79-3.67 (m, 6H), 3.63 (d, J = 5.7 Hz, 2H), 3.54 (dd, J = 4.9 Hz, J = 16.2 Hz, 1H), 3.18-3.13 (m, 2H), 3.04 (dd, J = 4.9 Hz, J = 13.9 Hz, 1H), 2.88 (dd, J = 8.6 Hz, J = 13.6 Hz, 1H), 2.04 (s, 3H), 1.96 (sep, J = 6.8 Hz, 2H), 1.87 (s, 3H), 1.80–1.75 (m, 1H), 1.56–1.47 (m, 3H), 0.87–0.82 (m, 12H). **g**, ^{1}H-NMR of CR peptide (Ac-NH-VVCGGGRGG-C(O)NH_2). ^{1}H-NMR (500 MHz, d_6-DMSO) δ 8.27–8.24 (m, 2H), 8.18 (t, J = 5.7 Hz, 1H), 8.13–8.08 (m, 3H), 8.04 (t, J = 5.7 Hz, 1H), 7.91 (d, J = 8.8 Hz), 7.86 (d, J = 8.8 Hz, 1H), 7.43 (t, J = 5.4 Hz, 1H), 7.28 (s, 1H), 7.10 (s, 1H), 4.39 (dt, J = 5.6 Hz, J = 7.4 Hz, 1H), 4.28 (dt, J = 5.7 Hz, J = 7.2 Hz, 1H), 4.21-4.13 (m, 2H), 3.82-3.70 (m, 8H), 3.64 (d, J = 5.8, 2H), 3.08 (dt, J = 6.5 Hz, J = 6.5 Hz, 2H), 2.80-2.67 (m, 2H), 2.43 (t, J = 8.6 Hz, 1H), 1.94 (sep, J = 6.8 Hz, 2H), 1.85 (s, 3H), 1.75–1.68 (m, 1H), 1.54–1.42 (m, 3H), 0.85–0.81 (m, 12H) **h**, ^{1}H–^{1}H total correlation spectroscopy (TOCSY) of CR-MGO peptide. **i**, Peak assignment for CR-MGO peptide TOCSY spectrum. Data are mean ± s.e.m. of biologically independent samples.



Extended Data Fig. 8 | MS2 analysis of CR-MGO-crosslinked KEAP1 peptide. a, Targeted PRM transitions ($n = 6$). **b**, Annotated MS/MS spectrum from the crosslinked C151–R135 KEAP1 peptide.

Extended Data Table 1 | Primer sequences for qPCR and cloning experiments

Gene	Forward Primer Sequence	Reverse Primer Sequence
NQO1	GCCTCCTTCATGGCATAGTT	GGACTGCACCAGAGCCAT
HMOX1	GAGTGTAAGGACCCATCGGA	GCCAGCAACAAAGTGCAAG
ME1	GGAGACGAAATGCATTCACA	ACGAATTCATGGAGGCAGTT
GCLM	GCTTCTTGGAAACTTGCTTCA	CTGTGTGATGCCACCAGATT
TXNRD1	TCAGGGCCGTTCATTTTTAG	GATCTGCCCGTTGTGTTTG
FTH1	GGCAAAGTTCTTCAAAGCCA	CATCAACCGCCAGATCAAC
GSR	TTGGAAAGCCATAATCAGCA	CAAGCTGGGTGGCACTTG
EPHX1	CTTCACGTGGATGAAGTGGA	CTGGCGGAATGAATTTGACT
ABCC2	GGGATCTCTTCCACACTGGAT	CATACAGGCCCTGAAGAGGA
PRDX1	GGGCACACAAAGGTGAAGTC	GCTGTTATGCCAGATGGTCAG
NQO2	TGCGTAGTCTCTCTTCAGCG	GCAACTCCTAGAGCGGTCCT
GSTM3	GGGTGATCTTGTTCTTCCCA	GGGGAAGCTCCTGACTATGA
SOD1	CCACACCTTCACTGGTCCAT	CTAGCGAGTTATGGCGACG
TXNRD1	TCAGGGCCGTTCATTTTTAG	GATCTGCCCGTTGTGTTTG
GSTP1	CTCAAAAGGCTTCAGTTGCC	ACCTCCGCTGCAAATACATC
GCLC	CTTTCTCCCCAGACAGGACC	CAAGGACGTTCTCAAGTGGG
GLO1	TGGATTAGCGTCATTCCAAG	GCGGACCCCAGTACCAAG
PGK1	CTTGGGACAGCAGCCTTAAT	CAAGCTGGACGTTAAAGGGA
TUBG1	ATCTGCCTCCCGGTCTATG	TACCTGTCGGAACATGGAGG

Mutation	Primer (Forward)	Primer (Reverse)
C23S	5'-/5Phos/GCA GGG GAC GCG GTG ATG TAC -3'	5'-/5Phos/CCC CTC AGG AGA CTG TGA CTG CAG GGG C -3'
C38S	5'-/5Phos/GCC CTC CCA GCA TGG CAA -3'	5'-/5Phos/GTC ACC TCC GCC TTG GAC TCA GT -3'
C151S	5'-/5Phos/TGA ACG GTG CTG TCA TGT ACC AGA TC -3'	5'-/5Phos/TGA CGT GGA GGA CAG ACT TCT CGC -3'
C273S	5'-/5Phos/CCG AAC TTC CTG CAG ATG CAG CT -3'	5'-/5Phos/CGT CAA CGA GTG GGA GCG CAC G -3'
C288S	5'-/5Phos/GTC CGA CTC CCG CTG CAA GGA CT -3'	5'-/5Phos/TGC AGG ATC TCG GAC TTC TGC AGC T -3'
C396S	5'-/5Phos/GAC CAA TCA GTG GTC GCC CTG -3'	5'-/5Phos/ATG GGG TTG TAA GAG TCC AGG GC -3'
C405S	5'-/5Phos/CGT GCC CCG TAA CCG CAT CG -3'	5'-/5Phos/CTC ATG GGG GCG CTG GGC G -3'
K39R	5'-/5Phos/GCC CTC CCA GCA TGG CAA -3'	5'-/5Phos/GTC ACC TCC GCC CTG CAC TCA GT -3'
K39M	5'-/5Phos/GCC CTC CCA GCA TGG CAA -3'	5'- GTC ACC TCC GCC ATG CAC TCA GT -3'
C38S/K39M	5'-/5Phos/GCC CTC CCA GCA TGG CAA -3'	5'- GTC ACC TCC GCC ATG GAC TCA GT -3'
K150M	5'-/5Phos/TGA ACG GTG CTG TCA TGT ACC AGA TC -3'	5'- TGA CGT GGA GGA CAC ACA TCT CGC C -3'
R6A	5'- GCA GCC AGA TCC CGC GCC TAG CGG GGC TG -3'	5'- CAG CCC CGC TAG GCG CGG GAT CTG GCT GC -3'
R15A	5'- GGG CCT GCT GCG CAT TCC TGC CCC TGC A -3'	5'- TGC AGG GGC AGG AAT GCG CAG CAG GCC C -3'
R50A	5'- CTC CCA GCA TGG CAA CGC CAC CTT CAG CTA CAC -3'	5'- GTG TAG CTG AAG GTG GCG TTG CCA TGC TGG GAG -3'
R135A	5'- CCC AAG GTC ATG GAG GCC CTC ATT GAA TTC GCC T -3'	5'- AGG CGA ATT CAA TGA GGG CCT CCA TGA CCT TGG G -3'

Extended Data Table 2 | Acquisition parameters used for targeted metabolomic measurements on a triple quadrupole mass spectrometer

Metabolite	Precursor mass	MS1 Resolution	Product ion	MS2 Resolution	Dwell	Fragmentor	Collision Energy	Polarity	Retention time (min)
Glucose	179.05	Wide	89.2	Unit	5	68	12	Neg	12.2
G6P	258.9	Wide	138.9	Unit	100	100	5	Neg	22.3
FBP	339.1	Wide	96.9	Unit	100	100	20	Neg	26.8
GAP	169	Wide	96.9	Unit	100	100	5	Neg	22.1
BPG	264.9	Wide	96.9	Unit	5	86	21	Neg	30.9
2/3-PG	184.98	Wide	78.9	Unit	5	86	21	Neg	24.6
PEP	166.97	Wide	79	Unit	5	78	9	Neg	25.4
Pyruvate	87.1	Wide	43	Unit	100	100	10	Neg	14.8
Lac	89.1	Wide	43	Unit	100	100	20	Neg	13.5
D3-Serine	107.05	Wide	75.1	Unit	5	18	9	Neg	13.9
R5P	228.7	Wide	78.8	Unit	100	100	35	Neg	19.9
Serine	104.2	Wide	73.8	Unit	5	100	5	Neg	13.9
GSH	305.7	Wide	143.0	Unit	100	100	15	Neg	16.7
GSSG	610.7	Wide	305.9	Unit	100	100	15	Neg	20.5
Succ	117	Wide	73.1	Unit	100	100	5	Neg	18.8
Glu	146.1	Wide	102.1	Unit	100	100	5	Neg	15.9
Cit	191	Wide	111	Unit	5	100	5	Neg	24.4
NAD^+	662.1	Wide	540	Unit	100	100	15	Neg	16.1
NADH	663.4	Wide	407.9	Unit	100	100	35	Neg	16.1
$NADP^+$	742	Wide	619.9	Unit	100	100	25	Neg	24.1
NADPH	743.5	Wide	407.8	Unit	100	100	25	Neg	24.1
ATP	506	Wide	159	Unit	100	100	25	Neg	27.5
ADP	425.8	Wide	134	Unit	100	100	15	Neg	26.5
3PGha	199.98	Wide	199.98	Unit	5	116	0	Neg	22.4
3PGha	199.98	Wide	79	Unit	5	116	15	Neg	22.4
2MQ	145.1	Wide	77.1	Unit	5	100	24	Pos	8.5
2MQ	145.1	Wide	92.1	Unit	5	100	20	Pos	8.5
D3-Serine	109.07	Wide	63.1	Unit	5	40	12	Pos	4.3

2/3-PG, 2/3-phosphoglycerate; 2MQ, 2-methylquinoxaline(derivatization product of MGO); 3PGha, 3-phosphoglyceroyl hydroxamic acid (derivatization product of 1,3-BPG). d_3-serine is an isotopically labelled serine standard included in all runs as an internal normalization control.

https://doi.org/10.1038/s41586-018-0615-z

LILRB4 signalling in leukaemia cells mediates T cell suppression and tumour infiltration

Mi Deng[1,20], Xun Gui[2,20], Jaehyup Kim[3,20], Li Xie[4], Weina Chen[3], Zunling Li[1,5], Licai He[1,6], Yuanzhi Chen[2,7], Heyu Chen[1], Weiguang Luo[1,8], Zhigang Lu[1,9], Jingjing Xie[1,5], Hywyn Churchill[3], Yixiang Xu[2], Zhan Zhou[1], Guojin Wu[1], Chenyi Yu[2,10], Samuel John[11], Kouyuki Hirayasu[12], Nam Nguyen[1], Xiaoye Liu[1], Fangfang Huang[1,13], Leike Li[2], Hui Deng[2], Haidong Tang[3], Ali H. Sadek[1], Lingbo Zhang[1,10], Tao Huang[14], Yizhou Zou[8], Benjamin Chen[15], Hong Zhu[16,17], Hisashi Arase[12], Ningshao Xia[7], Youxing Jiang[1], Robert Collins[18], M. James You[19], Jade Homsi[18], Nisha Unni[18], Cheryl Lewis[17], Guo-Qiang Chen[4], Yang-Xin Fu[3], X. Charlene Liao[14], Zhiqiang An[2]*, Junke Zheng[4]*, Ningyan Zhang[2]* & Cheng Cheng Zhang[1]*

Immune checkpoint blockade therapy has been successful in treating some types of cancer but has not shown clinical benefits for treating leukaemia[1]. This result suggests that leukaemia uses unique mechanisms to evade this therapy. Certain immune inhibitory receptors that are expressed by normal immune cells are also present on leukaemia cells. Whether these receptors can initiate immune-related primary signalling in tumour cells remains unknown. Here we use mouse models and human cells to show that LILRB4, an immunoreceptor tyrosine-based inhibition motif-containing receptor and a marker of monocytic leukaemia, supports tumour cell infiltration into tissues and suppresses T cell activity via a signalling pathway that involves APOE, LILRB4, SHP-2, uPAR and ARG1 in acute myeloid leukaemia (AML) cells. Deletion of *LILRB4* or the use of antibodies to block LILRB4 signalling impeded AML development. Thus, LILRB4 orchestrates tumour invasion pathways in monocytic leukaemia cells by creating an immunosuppressive microenvironment. LILRB4 represents a compelling target for the treatment of monocytic AML.

To identify novel mechanisms for AML development and immune regulation, we analysed the relationship between gene expression of known co-stimulating and co-inhibitory receptors and the overall survival of patients with AML as documented in The Cancer Genome Atlas (TCGA) database. Expression of the mRNA that encodes leukocyte immunoglobulin-like receptor B4 (LILRB4*)*, an immune inhibitory receptor that is restrictively expressed on monocytic cells[2–4] and monocytic AML cells (FAB M4 and M5 AML subtypes)[5], ranked at the top of the list for negative correlation with patient survival (Fig. 1a, Extended Data Fig. 1a–d, Supplementary Table 1). Notably, LILRB4 levels were higher on monocytic AML cells than on normal monocytes (Fig. 1b).

The extracellular domain of LILRB4 inhibits T cell activity[6]. To test whether LILRB4 expressed on AML cells suppresses T cells, we cultured LILRB4-positive leukaemia cells, LILRB4-negative leukaemia cells, or normal haematopoietic cells together with either autologous T cells or T cells from healthy donors. Only LILRB4-positive monocytic AML cells substantially suppressed T cell proliferation (Fig. 1c, Extended Data Fig. 1e, f). We then deleted *LILRB4* from human monocytic AML THP-1 and MV4-11 cells and found that the ability of AML cells to suppress T cells was reduced upon *LILRB4* knockout (*LILRB4*^{KO}) and was restored by forced expression of wild-type *LILRB4* (*LILRB4*^{KO}-WT), but not by forced expression of a mutant *LILRB4* with deleted intracellular domain (as *LILRB4*^{KO}-intΔ) (Fig. 1d, Extended Data Fig. 2a–g). Moreover, when wild-type THP-1 cells and human T cells were cultured in separate transwells, LILRB4-mediated T cell inhibition was also observed and could be reversed by anti-LILRB4 blocking antibodies (Extended Data Fig. 2h–p). Blocking LILRB4 resulted in an increase in T cell cytotoxicity and cytokine release (Extended Data Fig. 2q, u). These in vitro data suggest that, instead of the extracellular domain[6], the intracellular signalling of LILRB4 in AML cells is required for suppression of T cell activity.

Next, we used humanized mouse xenograft models and an immunocompetent mouse model to investigate LILRB4 function in immune checkpoint blockade. Subcutaneous implantation of THP-1 cells—but not of *LILRB4*^{KO} THP-1 cells—resulted in the development of AML in human T cell-reconstituted mice, and this was blocked by anti-LILRB4 treatment[7] (Extended Data Fig. 3a–i). Doxycycline-induced deletion of *LILRB4* in an established disseminated leukaemia model in humanized mice also impaired leukaemia development and restored T cells (Fig. 1e, f, Extended Data Fig. 3j–l). In addition, we subcutaneously implanted human LILRB4-expressing mouse C1498 AML cells (h*LILRB4*-C1498) into C57BL/6 mice to establish a syngeneic immunocompetent mouse model. To exclude the anti-tumour effects of Fc effector functions, we treated tumour-bearing mice with anti-LILRB4 with the Fc glycosylation site N297A mutation[8]. Blockade of LILRB4 effectively lowered tumour burden and prolonged survival; depletion of CD8^{+} T cells eliminated the anti-tumour effects of the anti-LILRB4 antibody (Extended Data Fig. 3m–r). These results suggest that the tumour-supportive effect of LILRB4 depends on inhibition of host T cells. The anti-LILRB4 antibody treatment generated tumour-specific memory T cells (Extended Data Fig. 3s). Similar results were obtained in the disseminated h*LILRB4*-C1498 syngeneic mouse model (Extended Data Fig. 3x–z). Finally, blockade of LILRB4 reduced leukaemia development in xenografts derived from primary human monocytic AML cells (Fig. 1g–i, Extended Data Fig. 4a) and increased the number of engraftable autologous human T cells (Extended Data Fig. 4b). Together, our in vitro and in vivo results

[1]Department of Physiology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Texas Therapeutics Institute, Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center, Houston, TX, USA. [3]Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [4]Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of Chinese Ministry of Education, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [5]Taishan Immunology Program, Basic Medicine School, Binzhou Medical University, Yantai, China. [6]Key Laboratory of Laboratory Medicine, Ministry of Education, School of Laboratory Medical and Life Science, Wenzhou Medical University, Wenzhou, China. [7]School of Public Health, Xiamen University, Xiamen, China. [8]Department of Immunology, Xiangya Medical School, Central South University, Changsha, China. [9]Institute of Biomedical Sciences and the Fifth People's Hospital of Shanghai, Fudan University, Shanghai, China. [10]Xiangya Medical School, Central South University, Changsha, China. [11]Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [12]Department of Immunochemistry, Research Institute for Microbial Diseases and Laboratory of Immunochemistry, World Premier International Immunology Frontier Research Center, Osaka University, Osaka, Japan. [13]Department of Hematology, Zhongshan Hospital, Xiamen University, Xiamen, China. [14]Immune-Onc Therapeutics, Inc., Palo Alto, CA, USA. [15]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [16]Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA. [17]Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA. [18]Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, USA. [19]Department of Hematopathology, Division of Pathology and Laboratory Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [20]These authors contributed equally: Mi Deng, Xun Gui, Jaehyup Kim. *e-mail: zhiqiang.an@uth.tmc.edu; zhengjunke@sjtu.edu.cn; ningyan.zhang@uth.tmc.edu; alec.zhang@utsouthwestern.edu

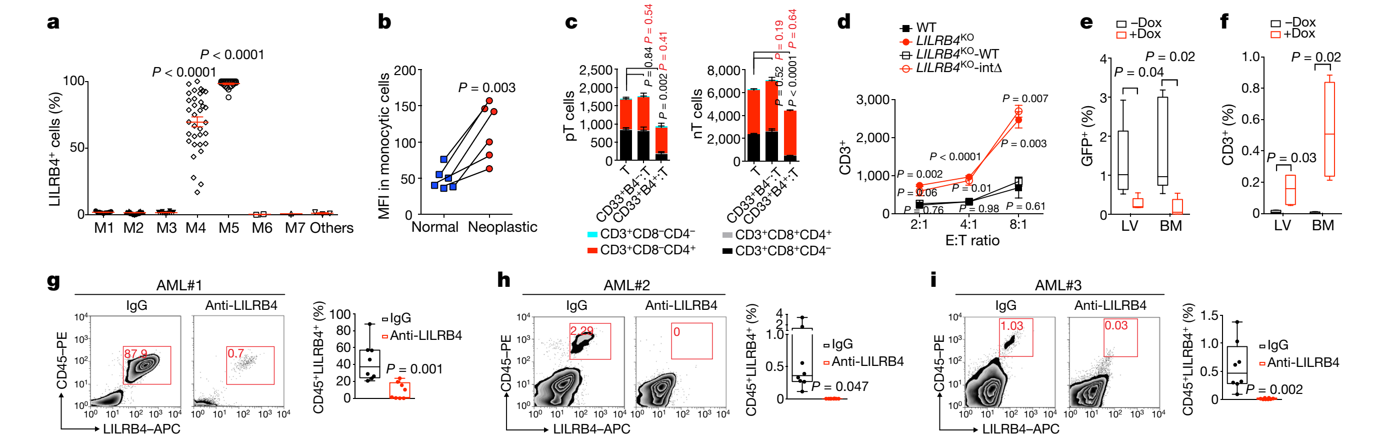


Fig. 1 | LILRB4 expressed on leukaemia cells suppresses T cell proliferation. a, Surface expression of LILRB4 quantified by flow cytometry analysis of samples from 105 patients with AML ($n = 1$–34 for each classification (see Methods), mean ± s.e.m.). **b**, Comparison of surface expression of LILRB4 on normal monocytes and neoplastic monocytes from the same patients ($n = 6$ independent patients). MFI, mean fluorescence intensity. **c**, Autologous T cells (pT, patient T cells) isolated from a patient with monocytic AML (AML#19) or allogeneic T cells (nT, normal T cells) isolated from a healthy donor were incubated with irradiated LILRB4$^+$ or LILRB4$^-$ (B4$^+$ or B4$^-$) primary leukaemia cells from patient AML#19 ($n = 3$ biologically independent samples, mean ± s.e.m.; see Source Data). **d**, T cells (E, effector cells) isolated from healthy donors were incubated with indicated wild-type or modified irradiated THP-1 cells (T, target cells) in a cell-contact manner ($n = 3$ biologically independent samples, mean ± s.e.m.). WT, wild-type. **e**, **f**, Engraftment of human T cells and intravenously transplanted doxycycline (Dox)-inducible *LILRB4*^{KO} THP-1 cells (GFP$^+$) into NOD–SCID IL2Rγ-null (NSG) mice ($n = 5$ mice). LV, liver; BM, bone marrow; %, percentage of leukaemia cells (GFP$^+$) or human T cells (CD3$^+$) in total mouse liver or bone marrow cells. **g**–**i**, Representative flow plots and quantification of per cent of CD45$^+$LILRB4$^+$ cells in bone marrow from mice xenografted with human primary monocytic AML cells after treatment with anti-LILRB4 antibody or control IgG ($n = 8$ biologically independent samples). Experiment repeated for 16 independent patient samples with similar results (Extended Data Fig. 4a). See Methods for definition of box plot elements in **e**–**i**. All *P* values from two-tailed Student's *t*-test.

indicate that LILRB4 signalling in monocytic AML cells suppresses T cell-mediated anti-tumour immunity.

One of the characteristic features of monocytic AML is enhanced extramedullary infiltration of tumour cells[9]. In addition to tumour shrinkage (Extended Data Fig. 3m), antibody blockade of LILRB4 resulted in a decrease in leukaemic infiltration into internal organs, including bone marrow, liver and brain (Extended Data Fig. 3t–v). Although anti-LILRB4 antibody treatment did not reduce the size

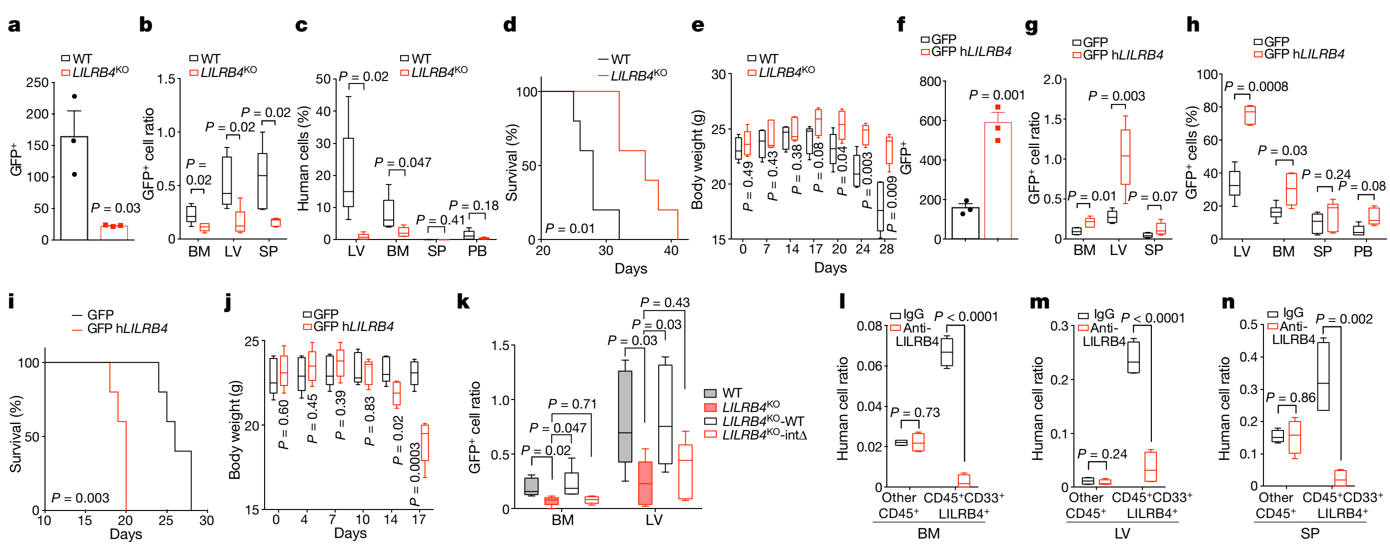


Fig. 2 | LILRB4 promotes migration and infiltration of AML cells. a, Comparison of the transendothelial migration abilities of wild-type and *LILRB4*^{KO} THP-1 cells (GFP$^+$) in a transwell assay ($n = 3$ biologically independent samples, mean ± s.e.m.). **b**, Comparison of the short-term (20 h) infiltration of wild-type or *LILRB4*^{KO} THP-1 cells in NSG mice ($n = 5$ mice). The numbers of leukaemia cells (GFP$^+$) in liver (LV), spleen (SP) and bone marrow (BM) were determined by flow cytometry and normalized to number in peripheral blood (PB). **c**–**e**, Comparison of long-term (21 days) infiltration of wild-type or *LILRB4*^{KO} THP-1 cells in NSG mice ($n = 5$ mice). **c**, Percentages of THP-1 cells (hCD45$^+$) engrafted in indicated organs at day 21 post-transplant. **d**, **e**, Overall survival (**d**) and body weight (**e**) over time. **f**–**j**, Comparison of transendothelial migration (**f**, $n = 3$ biologically independent samples, mean ± s.e.m.), short-term (**g**, 20 h, $n = 5$ mice) and long-term (**h**, 16 days, $n = 5$ mice) infiltration of h*LILRB4* C1498 or control C1498 cells (GFP$^+$), and overall survival (**i**) and body weight (**j**) over time. **k**, Comparison of short-term (20 h) infiltration of indicated wild-type or modified THP-1 cells in NSG mice ($n = 5$ mice). **l**–**n**, Comparison of short-term (20 h) infiltration of human primary monocytic AML cells (AML#21) in NSG mice ($n = 4$ mice) after treatment with anti-LILRB4 antibody or IgG control. See Methods for definition of box plot elements in **b**, **c**, **e**, **g**, **h**, **j**–**n**. All *P* values from two-tailed Student's *t*-test except for **d**, **i** from long-rank test.

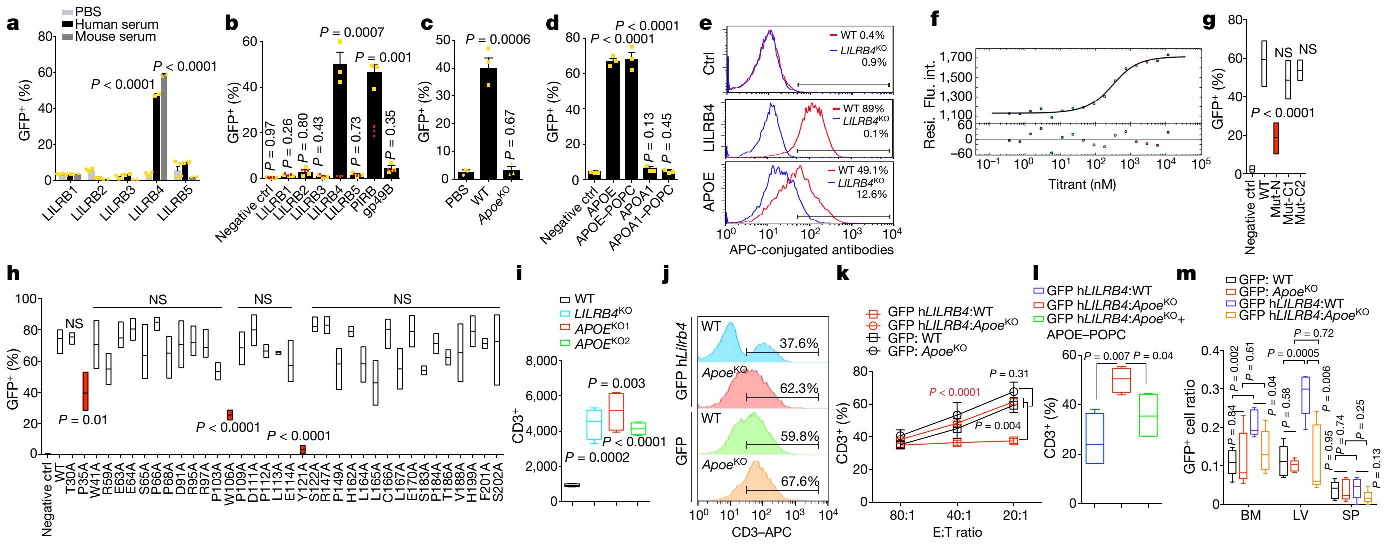


Fig. 3 | APOE is an extracellular binding protein of LILRB4. **a**, Percentages of indicated LILRB reporter cells activated (GFP$^+$) in the presence of 10% human serum, 10% mouse serum or PBS control. **b**, Percentages of indicated LILRB reporter cells activated by recombinant APOE (10 μg ml^{-1}). Red dots indicate PBS treatment of each indicated reporter cell line. **c**, Percentages of LILRB4 reporter cells activated by 10% mouse serum collected from wild-type or *Apoe* knockout (*Apoe*^{KO}) mice or PBS control. **d**, Percentages of LILRB4 reporter cells activated by 10 μg ml^{-1} APOE, APOE–POPC, APOA1 or APOA1–POPC. Mean ± s.e.m. (**a**–**d**). **e**, Binding of His-tagged APOE to wild-type and *LILRB4*^{KO} THP-1 cells. **f**, Binding kinetics of human His-tagged APOE3 to the extracellular domain of LILRB4 (LILRB4-ECD) were measured using MST. Top, fluorescence intensity (Flu. int.) plot and regression of the binding; bottom, corresponding residuals (Resi.) versus fits plot. **g**, Percentages of LILRB4 reporter cells activated by wild-type and mutant APOE proteins. Mut-N, R142A/K143A/R145A/K146A/R147A/R150A; Mut-C1, deletion of residues 245–299; Mut-C2, deletion of residues 279–299. **h**, Percentages of indicated LILRB4 mutant reporter cells activated by APOE proteins. Data on LILRB4 mutants that interfere with APOE activation are highlighted in red (**g**, **h**, low-to-high outline and line at mean in box plots). **i**, T cells from healthy donors were incubated with indicated irradiated wild-type, *LILRB4*^{KO} or *APOE*^{KO} THP-1 cells. T cells were analysed by flow cytometry after 7 days. **j**–**l**, C57BL/6 mouse spleen cells (E) were incubated with irradiated human *LILRB4*-expressing (GFP h*LILRB4*) or control (GFP) C1498 cells (T) at indicated E:T ratios. Cells were supplemented with 5% serum from wild-type or *Apoe*^{KO} mice, cultured with anti-CD3 and anti-CD28-coated beads for 60 h, and then stained with anti-CD3 antibody. **j**, Representative flow plots from samples at E:T of 20:1. **k**, Percentages of CD3$^+$ T cells (mean ± s.e.m.). **l**, Effects of APOE–POPC rescue of *Apoe*^{KO} serum. **m**, Expression of human *LILRB4* in mouse leukaemia C1498 cells increases leukaemia cell infiltration in wild-type recipient mice but not in *Apoe*^{KO} recipient mice. **e**, **j**, Experiments were repeated independently three times with similar results. See Methods for definition of box plot elements in **i**, **l**, **m**. All *P* values from two-tailed Student's *t*-test. NS, not significant. **a**–**d**, $n = 3$ biologically independent samples; **g**–**i**, **k**, **l**, $n = 4$ biologically independent samples; **m**, $n = 5$ mice.

of subcutaneous C1498 tumours in C57BL/6 mice depleted of CD8$^+$ T cells (Extended Data Fig. 3m), treatment with anti-LILRB4 antibody did lead to a decrease in leukaemia cell infiltration into the liver (Extended Data Fig. 3w). We hypothesized that, as well as inhibiting T cells, LILRB4 promotes leukaemia infiltration. To test this hypothesis, we performed transendothelial migration and homing assays and monitored leukaemia infiltration relative to LILRB4 expression on leukaemia cells. Human AML THP-1 cells depleted of LILRB4 had lower transendothelial migration in vitro than cells that expressed LILRB4 (Fig. 2a). Deletion of *LILRB4* reduced homing and engraftment of AML cells to haematopoietic organs (Fig. 2b, c), and resulted in prolonged survival of xenografted mice (Fig. 2d) and delayed body weight loss (Fig. 2e). By contrast, forced expression of human LILRB4 in mouse AML C1498 or WEHI-3 cells had the opposite effects (Fig. 2f–j, Extended Data Fig. 5a–e). Antibody-mediated LILRB4 blockade in LILRB4-expressing AML cells had the same effects as LILRB4 knockout (Extended Data Fig. 5f–t). Leukaemia infiltration depended on LILRB4 expression and its intracellular signalling in leukaemia cells (Fig. 2k) but not the Fc effector functions of the antibody (Extended Data Fig. 5u, v). Furthermore, LILRB4 blockade reduced the infiltration ability of primary monocytic AML cells (Fig. 2l–n, Extended Data Fig. 4c–e). Our results are concordant with previous studies showing that the frequency of circulating LILRB4$^+$ AML blasts is lower than that of LILRB4$^-$ AML blasts[5] and that LILRB4$^+$ chronic lymphocytic leukaemia cells are associated with lymphoid tissue involvement[10]. The bone marrow, liver, and brain—to which LILRB4$^+$ AML cells tend to migrate—are known to have certain immune privileges[11–13]. Thus, LILRB4-mediated migration, which supports enhanced extramedullary infiltration of monocytic AML cells, may also contribute to immune evasion.

The blockade of immune inhibitory and migration functions of AML cells by anti-LILRB4 antibodies suggests that these functions of LILRB4 are regulated by extracellular mechanisms. Integrin-$\alpha_v\beta_3$ is the ligand for gp49B1, the mouse LILRB4 orthologue[14]. However, a variety of integrin-αβ complexes did not activate human LILRB4 reporter cells (Extended Data Fig. 6a, b). Unexpectedly, human serum and mouse serum were capable of activating the LILRB4 reporter but not reporters for other LILRBs (Fig. 3a). Through protein liquid chromatography fractionation followed by reporter assays and mass spectrometry we identified APOE, which specifically activated LILRB4 and mouse PirB reporters (Fig. 3b, Extended Data Fig. 6c–j). Serum from wild-type mice, but not *Apoe*-null mice, activated the LILRB4 reporter (Fig. 3c). In addition, both liposome-reconstituted APOE protein (APOE–POPC) and lipid-free APOE activated LILRB4 reporter cells (Fig. 3d). Binding of APOE to THP-1 cells was significantly decreased by deletion of *LILRB4* (Fig. 3e). We confirmed specific binding of recombinant APOE to LILRB4 using microscale thermophoresis (MST), surface plasmon resonance (SPR) and bio-layer interferometry (Octet). The dissociation constant was 210 nM as determined by MST (Fig. 3f, Extended Data Fig. 6k, l). Mutagenesis studies showed that the N-terminal domain of APOE, and P35 and W106 in the first immunoglobulin domain and Y121 in the linker region between two immunoglobulin domains of LILRB4, are critical for APOE-mediated activation of LILRB4 (Fig. 3g, h, Supplementary Table 2, Extended Data Fig. 6m).

The finding that APOE activates the immune inhibitory receptor LILRB4 is consistent with the immunosuppressive function of APOE[15,16]. To determine whether suppression of T cells by LILRB4 depends on APOE, we examined proliferation of T cells co-cultured

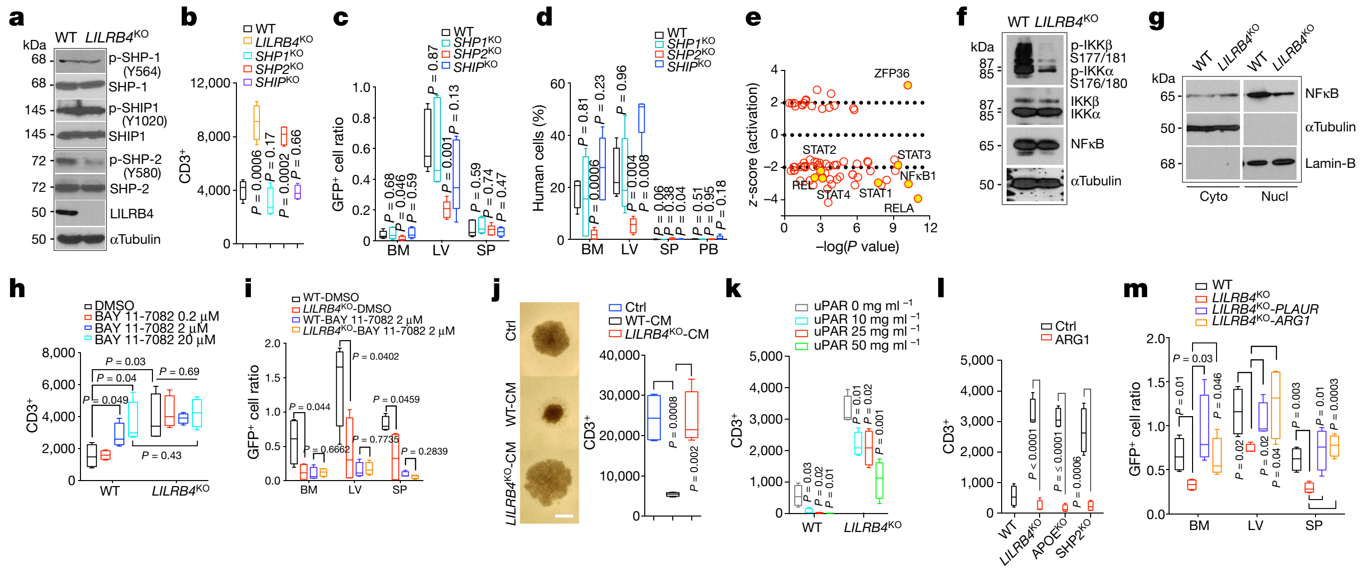


Fig. 4 | LILRB4-mediated intracellular signalling controls AML cell migration and T cell suppression. a, Expression and phosphorylation of three phosphatases in wild-type and *LILRB4*^{KO} THP-1 cells. **b**, Primary T cells and irradiated indicated THP-1 cells were cultured in the lower and upper chambers, respectively. T cells were analysed by flow cytometry after 7 days. $n = 4$ biologically independent samples. **c**, **d**, Knockout of *SHP2* reduces short-term (20 h) and long-term (21 days) infiltration of THP-1 cells in NSG mice ($n = 5$ mice). **e**, Upstream transcription factor analysis of RNA sequencing (RNA-seq) data generated from *LILRB4*^{KO} and wild-type THP-1 cells ($n = 2$ biologically independent samples). Yellow dots highlight the transcription factors involved in the JAK–STAT and NFκB pathways. **f**, Decreased phosphorylation of IKKα and IKKβ in *LILRB4*^{KO} THP-1 cells. **g**, Decreased NFκB in the nuclear fraction in *LILRB4*-KO THP-1 cells. **h**, **i**, An NFκB inhibitor (BAY 11-7082) reversed T cell suppression by THP-1 cells (**h**) and decreased infiltration of MV4-11 cells (**i**) in an LILRB4-dependent manner ($n = 4$ biologically independent samples). **j**, T cells isolated from healthy donors were supplemented with 25% conditioned medium (CM) from wild-type or *LILRB4*^{KO} THP-1 cells. Left, representative cells (scale bar, 100 μm); right, T cells were analysed by flow cytometry ($n = 4$ biologically independent samples). **k**, **l**, T cells were incubated with irradiated indicated THP-1 cells supplemented with indicated concentrations of recombinant uPAR (**k**) or ARG1 (**l**) proteins for 7 days and were analysed by flow cytometry ($n = 4$ biologically independent samples). **m**, Overexpression of uPAR (*PLAUR*) or *ARG1* rescued the infiltration defect in *LILRB4*^{KO} MV4-11 cells ($n = 5$ mice). **a**, **f**, **g**, **j**, Experiments repeated independently three times with similar results. See Methods for definition of box plot elements in **b–d**, **h–m**. All *P* values from two-tailed Student's *t*-test.

with control or *APOE*^{KO} human AML cells. AML cells deficient in APOE restored proliferation of T cells and suppressed migration of leukaemia cells (Fig. 3i, Extended Data Fig. 6n–t). Moreover, the percentage of T cells in co-culture was lower when C1498 cells ectopically expressing LILRB4 were treated with wild-type mouse serum compared to those treated with *Apoe*^{KO} mouse serum (Fig. 3j, k). Addition of liposome-reconstituted APOE to a co-culture of mouse spleen cells and LILRB4-expressing AML cells decreased the T cell percentage (Fig. 3l). Furthermore, expression of LILRB4 increased infiltration of C1498 cells to bone marrow and liver in wild-type mice but not in *Apoe*^{KO} recipients (Fig. 3m). These data indicate that APOE activates LILRB4 on human monocytic AML cells to suppress T cell proliferation and support AML cell migration.

We sought to identify the signalling downstream of LILRB4 that is required for T cell suppression and leukaemia infiltration. The phosphatases SHP-1 (also known as PTPN6), SHP-2 (also known as PTPN11) and SHIP (also known as INPP5D) can be recruited to the intracellular domain of LILRB[2]. The level of phosphorylation of SHP-2, but not of SHP-1 or SHIP, was lower in *LILRB4*^{KO} AML cells than in wild-type cells (Fig. 4a, Extended Data Fig. 7a). Loss of SHP-2, but not loss of SHP-1 or SHIP, reversed T cell suppression by THP-1 cells (Fig. 4b, Extended Data Fig. 7b–d), and decreased short-term (20 h) and long-term (21 days) infiltration of THP-1 cells (Fig. 4c, d). Our results suggest that SHP-2 is a mediator of LILRB4 signalling.

Our ingenuity pathway analysis showed that the activity of the key transcription factors NFκB1 and RELA in the NFκB pathway[17], which is positively regulated by SHP-2[18], was strongly inhibited by loss of *LILRB4* (Fig. 4e, Supplementary Tables 3, 4). Consistently, phosphorylation of IKKα/β and levels of nuclear NFκB were decreased in *LILRB4*^{KO} AML cells (Fig. 4f, g, Extended Data Fig. 7a). Inhibition of NFκB signalling reversed T cell suppression and reduced AML cell infiltration in a LILRB4-dependent manner (Fig. 4h, i, Extended Data Fig. 7e, f). Therefore the effects of LILRB4 are mediated through the NFκB pathway, which is particularly robust in monocytic AML among AML subtypes[19].

Consistent with our result that AML cells inhibit T cell proliferation in transwells (Extended Data Fig. 2o, p), conditioned medium from wild-type THP-1 cells suppressed T cell activity but conditioned medium from *LILRB4*^{KO} cells did not (Fig. 4j). Among proteins that were present at higher levels in the conditioned medium of wild-type THP-1 cells than in that of *LILRB4*^{KO} cells (Extended Data Fig. 7g–i), the urokinase receptor uPAR is highly expressed by monocytic AML cells[20]. uPAR, an NFκB target, is known to promote cancer invasion, metastasis, survival and angiogenesis[21,22]. The addition of recombinant uPAR decreased proliferation of T cells co-cultured with *LILRB4*^{KO} THP-1 cells in a dose-dependent manner (Fig. 4k, Extended Data Fig. 7j). This activity of uPAR is likely to be mediated by downstream effectors in AML cells because uPAR does not effectively decrease T cell proliferation directly (Extended Data Fig. 7k).

The expression of arginase-1 (ARG1), as with that of uPAR, was lower in *LILRB4*^{KO} AML cells than in wild-type cells (Extended Data Fig. 7l, m). ARG1 inhibits T cell proliferation and can be upregulated by uPAR-mediated signalling[23,24], APOE[25] and NFκB[26]. We hypothesized that ARG1 is a key downstream effector of LILRB4–NFκB–uPAR signalling. ARG1 can be secreted by AML cells to inhibit T cell activity[27]. Recombinant ARG1 decreased T cell proliferation in co-culture with *LILRB4*^{KO}, *APOE*^{KO} and *SHP2*^{KO} AML or primary AML cells (Fig. 4l, Extended Data Fig. 7n–p). Moreover, addition and overexpression of either uPAR or ARG1 rescued the migration ability of *LILRB4*^{KO} AML cells in vitro and in vivo, respectively (Extended Data Fig. 7q, Fig. 4m). Together, our results indicate that LILRB4–SHP-2–NFκB–uPAR–ARG1 is a signalling pathway in monocytic AML cells (Extended Data Figs. 8, 9) that suppresses immune activity and supports leukaemia migration.

Because LILRB4 is restrictively expressed on normal monocytic cells[2], in which LILRB4 signalling may differ from that in leukaemia cells (Extended Data Fig. 9), and LILRB4 blockade did not significantly interfere with normal haematopoietic function (Extended Data Fig. 10), LILRB4 targeting may have minimal toxicity. Notably, LILRB4 is also expressed on certain other types of cancer, myeloid-derived suppressor cells, tolerogenic dendritic cells and tumour-associated macrophages[2,5,28–30]. Targeting LILRB4 may thus enable a combination of immunotherapy and targeted therapy in cancer treatment.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0615-z.

Received: 15 May 2016; Accepted: 15 August 2018;
Published online 17 October 2018.

1. Curran, E. K., Godfrey, J. & Kline, J. Mechanisms of immune tolerance in leukemia and lymphoma. *Trends Immunol*. **38**, 513–525 (2017).
2. Kang, X. et al. Inhibitory leukocyte immunoglobulin-like receptors: Immune checkpoint proteins and tumor sustaining factors. *Cell Cycle* **15**, 25–40 (2016).
3. Hirayasu, K. & Arase, H. Functional and genetic diversity of leukocyte immunoglobulin-like receptor and implication for disease associations. *J. Hum. Genet*. **60**, 703–708 (2015).
4. Trowsdale, J., Jones, D. C., Barrow, A. D. & Traherne, J. A. Surveillance of cell and tissue perturbation by receptors in the LRC. *Immunol. Rev*. **267**, 117–136 (2015).
5. Dobrowolska, H. et al. Expression of immune inhibitory receptor ILT3 in acute myeloid leukemia with monocytic differentiation. *Cytometry B Clin. Cytom*. **84**, 21–29 (2013).
6. Vlad, G. et al. Membrane and soluble ILT3 are critical to the generation of T suppressor cells and induction of immunological tolerance. *Int. Rev. Immunol*. **29**, 119–132 (2010).
7. Mosier, D. E., Gulizia, R. J., Baird, S. M. & Wilson, D. B. Transfer of a functional human immune system to mice with severe combined immunodeficiency. *Nature* **335**, 256–259 (1988).
8. Ha, S. et al. Isolation and characterization of IgG1 with asymmetrical Fc glycosylation. *Glycobiology* **21**, 1087–1096 (2011).
9. Straus, D. J. et al. The acute monocytic leukemias: multidisciplinary studies in 45 patients. *Medicine (Baltimore)* **59**, 409–425 (1980).
10. Colovai, A. I. et al. Expression of inhibitory receptor ILT3 on neoplastic B cells is associated with lymphoid tissue involvement in chronic lymphocytic leukemia. *Cytometry B Clin. Cytom*. **72**, 354–362 (2007).
11. Crispe, I. N. et al. Cellular and molecular mechanisms of liver tolerance. *Immunol. Rev*. **213**, 101–118 (2006).
12. Carson, M. J., Doose, J. M., Melchior, B., Schmid, C. D. & Ploix, C. C. CNS immune privilege: hiding in plain sight. *Immunol. Rev*. **213**, 48–65 (2006).
13. Fujisaki, J. et al. *In vivo* imaging of Treg cells providing immune privilege to the haematopoietic stem-cell niche. *Nature* **474**, 216–219 (2011).
14. Castells, M. C. et al. gp49B1-$\alpha_v\beta_3$ interaction inhibits antigen-induced mast cell activation. *Nat. Immunol*. **2**, 436–442 (2001).
15. Grainger, D. J., Reckless, J. & McKilligin, E. Apolipoprotein E modulates clearance of apoptotic bodies *in vitro* and *in vivo*, resulting in a systemic proinflammatory state in apolipoprotein E-deficient mice. *J. Immunol*. **173**, 6366–6375 (2004).
16. Ali, K., Middleton, M., Puré, E. & Rader, D. J. Apolipoprotein E suppresses the type I inflammatory response *in vivo*. *Circ. Res*. **97**, 922–927 (2005).
17. DiDonato, J. A., Mercurio, F. & Karin, M. NF-κB and the link between inflammation and cancer. *Immunol. Rev*. **246**, 379–400 (2012).
18. You, M., Flick, L. M., Yu, D. & Feng, G. S. Modulation of the nuclear factor κb pathway by Shp-2 tyrosine phosphatase in mediating the induction of interleukin (IL)-6 by IL-1 or tumor necrosis factor. *J. Exp. Med*. **193**, 101–110 (2001).
19. Baumgartner, B. et al. Increased IκB kinase activity is associated with activated NF-κB in acute myeloid blasts. *Leukemia* **16**, 2062–2071 (2002).
20. Béné, M. C. et al. CD87 (urokinase-type plasminogen activator receptor), function and pathology in hematological disorders: a review. *Leukemia* **18**, 394–400 (2004).
21. Su, S. C., Lin, C. W., Yang, W. E., Fan, W. L. & Yang, S. F. The urokinase-type plasminogen activator (uPA) system as a biomarker and therapeutic target in human malignancies. *Expert Opin. Ther. Targets* **20**, 551–566 (2016).
22. Wang, Y. et al. Identification of a novel nuclear factor-kappaB sequence involved in expression of urokinase-type plasminogen activator receptor. *Eur. J. Biochem*. **267**, 3248–3254 (2000).
23. Hu, J. et al. uPAR induces expression of transforming growth factor β and interleukin-4 in cancer cells to promote tumor-permissive conditioning of macrophages. *Am. J. Pathol*. **184**, 3384–3393 (2014).
24. Ilkovitch, D. & Lopez, D. M. Urokinase-mediated recruitment of myeloid-derived suppressor cells and their suppressive mechanisms are blocked by MUC1/sec. *Blood* **113**, 4729–4739 (2009).
25. Baitsch, D. et al. Apolipoprotein E induces antiinflammatory phenotype in macrophages. *Arterioscler. Thromb. Vasc. Biol*. **31**, 1160–1168 (2011).
26. Hagemann, T. et al. "Re-educating" tumor-associated macrophages by targeting NF-κB. *J. Exp. Med*. **205**, 1261–1268 (2008).
27. Mussai, F. et al. Acute myeloid leukemia creates an arginase-dependent immunosuppressive microenvironment. *Blood* **122**, 749–758 (2013).
28. de Goeje, P. L. et al. Immunoglobulin-like transcript 3 is expressed by myeloid-derived suppressor cells and correlates with survival in patients with non-small cell lung cancer. *OncoImmunology* **4**, e1014242 (2015).
29. Chang, C. C. et al. Tolerization of dendritic cells by T(S) cells: the crucial role of inhibitory receptors ILT3 and ILT4. *Nat. Immunol*. **3**, 237–243 (2002).
30. Suciu-Foca, N. et al. Soluble Ig-like transcript 3 inhibits tumor allograft rejection in humanized SCID mice and T cell responses in cancer patients. *J. Immunol*. **178**, 7432–7441 (2007).

Acknowledgements We thank the National Cancer Institute (1R01CA172268 and 5P30CA142543), the Leukemia & Lymphoma Society (1024-14 and TRP-6024-14), the March of Dimes Foundation (1-FY14-201), the Cancer Prevention and Research Institute of Texas (RP140402, DP150056, RP180435, PR150551, and RR150072), the Robert A. Welch Foundation (I-1834 and AU-0042-20030616), the National Natural Science Foundation of China (81570093, 81422001, and 81721004), the National Basic Research Program of China (2014CB965000), and the China Scholarship Council (201608330307) for support. We also thank G. Salazar for editing the manuscript and Y. Dang for RStudio coding.

Author contributions M.D. and C.C.Z. designed the study and wrote the manuscript. M.D., C.C.Z., X.G., N.Z., Z.A. and X.C.L. contributed to the experimental plan and data interpretation. M.D., Z.Li. and L.H. performed mouse experiments. M.D., X.G., L.X., Z.Li., Y.C., Z.Lu., Y.X., Z.Z., C.Y., L.L., H.D., Z.A., J.Z. and N.Z. performed antibody characterizations. M.D., X.G. and L.L. measured APOE–LILRB4 binding affinity. K.H., H.A., M.D., J.K., L.H. and J.X. performed reporter assays. W.C., H.C., R.C., M.J.Y., J.H., N.U. and C.L. provided primary patient samples. M.D., W.C., L.H. and H.C. performed flow cytometry analysis of primary patient cells. M.D., Z.Li. and H.C. performed CRISPR–Cas9 experiments. M.D., X.G., Z.Li., L.H., H.C., W.L. and G.W. performed plasmid constructions. M.D., Z.Li., L.H., H.C., W.L., J.X., S.J., X.L. and L.Z. performed in vitro T cell assays. N.N. and Y.J. produced lipid-bound APOE protein. M.D., Z.Li., L.H. and F.H. performed western blotting. H.T., A.H.S., T.H., Y.Z., B.C., N.X., G.-Q.C., Y.-X.F., X.C.L., Z.A., N.Z. and C.C.Z. helped with or advised on experiments and provided reagents. M.D. and H.Z. performed statistical analysis.

Competing interests The Board of Regents of the University of Texas System has filed patent applications with PCT Application Nos. PCT/US2016/020838, which covers anti-LILRB antibodies and their uses in detecting and treating cancer, and PCT/US2017/044171, which covers the methods for identifying LILRB-blocking antibodies. Authors C.C.Z., M.D., Z.A., N.Z., X.G. and J.Z. are listed as inventors of PCT/US2016/020838. Authors C.C.Z., Z.A., N.Z., M.D., J.K. and X.G. are listed as inventors of PCT/US2017/044171. Both patent applications have been exclusively licensed to Immune-Onc Therapeutics by the Board of Regents of the University of Texas System. Authors Z.A. and C.C.Z. are Scientific Advisory Board members with Immune-Onc Therapeutics, who also own equities and have a sponsored research agreement with Immune-Onc Therapeutics. Authors T.H. and X.C.L. are employees of and hold equities in Immune-Onc Therapeutics.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0615-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0615-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Z.A. or J.Z. or N.Z. or C.C.Z.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Mice. C57 BL/6J and NOD-SCID IL2Rγ-null (NSG) mice were purchased from and maintained at the animal core facility of University of Texas Southwestern Medical Center (UTSW). *Apoe*^{KO} (*Apoe*^{tm1Unc}) mice[31] were purchased from the Jackson Laboratory. Animal work described in this manuscript has been approved and conducted under the oversight of the UT Southwestern Institutional Animal Care and Use Committee (IACUC). For each experiment, the same sex- and age-matched (4–8 weeks) mice were used and randomly allocated to each group; and for tumour size measurement and in vivo lumina imaging experiments, experimenters were blinded to the treatment conditions of the mice. The minimum number of mice in each group was calculated based on results from our prior relevant studies[32–36]. For the subcutaneous tumour model, the tumour size was calculated as (width × width × length) cm^3. The maximal tumour measurement permitted by UTSW IACUC is 2 cm in diameter. In none of the experiments were these limits exceeded (see Source Data). We complied with all relevant ethical regulations and used approved animal study protocols.

Cell culture. 293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS) at 37 °C in 5% CO2 and normal O_2. Human umbilical vein/vascular endothelium cells (HUVECs) (ATCC, CRL-1730) were cultured in endothelial cell growth medium plus growth factor, cytokines and supplements (EGM-BulletKit, Lonza) at 37 °C in 5% CO2 and normal O_2. Human monocytic AML cells (THP-1 (ATCC, TIB-202), MV4-11 (ATCC, CRL-9591), and U937 (ATCC, CRL-1593.2)) and mouse AML cells (WEHI-3 (ATCC, TIB-68)) were cultured in Roswell Park Memorial Institute (RPMI) 1640 supplemented with 10% FBS at 37 °C in 5% CO_2 and normal O_2. Mouse AML cells (C1498 (ATCC, TIB-49)) were cultured in DMEM supplemented with 10% FBS at 37 °C in 5% CO_2 and normal O_2. All cell lines were routinely tested using a mycoplasma-contamination kit (R&D Systems).

Primary human leukaemia cells. Primary human AML and B-cell acute lymphoblastic leukaemia (B-ALL) samples were obtained from the tissue banks at UTSW and University of Texas MD Anderson Cancer Center (MDACC). Informed consent was obtained under protocols reviewed and approved by the Institutional Review Board at UTSW and MDACC (IRB STU 122013-023 by UTSW and LAB10-0682 by MDACC). The UTSW cohort included 105 patients with AML representative of AML subtypes by the French-American-British (FAB) classification, acute myeloblastic leukaemia with minimal maturation (M1, $n = 9$), acute myeloblastic leukaemia with maturation (M2, $n = 34$), acute promyelocytic leukaemia (M3, $n = 10$), acute myelomonocytic leukaemia (M4, $n = 34$), acute monocytic leukaemia (M5, $n = 25$), acute erythroid leukaemia (M6, $n = 2$), and acute megakaryoblastic leukaemia (M7, $n = 1$) and patients with undifferentiated leukaemia (AUL; $n = 1$) and transient myeloproliferative disorder (TAM; $n = 2$) (Supplementary Table 1). Samples were frozen in FBS with 10% DMSO and stored in liquid nitrogen. Primary leukaemia samples used in patient-derived xenografts (PDXs), co-culture, short-term infiltration assay and western blotting were summarized in Supplementary Table 5. We have complied with all relevant ethical regulations with approved study protocols.

Human normal monocytes and macrophages. Human normal monocytes ($CD14^+$ cells) were isolated by the AutoMACS Pro Separation System (Miltenyi Biotech) from the mononuclear cells fraction of normal peripheral blood. In brief, buffy coat was purchased from Interstate Blood Bank and the mononuclear cell layer was separated by Ficoll Hypaque (17144003, GE Lifesciences) density gradient separation. Mononuclear cells were treated with red blood cell lysis buffer to remove red blood cells and then incubated with CD14 microbead-conjugated antibody (130-050-201, Miltenyi Biotech) for 15 min at 4 °C. CD14-positive cells were then isolated using the positive selection program according to the manufacturer's protocol. One million $CD14^+$ cells were plated in macrophage culture medium, Iscove's modified Dulbecco's medium (IMDM) (12440053, Thermo Fisher) supplemented with 10% human AB serum (MT35060CI, Fisher Scientific), 1% NEAA (11-140-050, Fisher), 2 μM L-alanine-L-glutamine (SH3003402, Fisher), per each well of a 6-well plate and cultured for 7 days. After incubation, most of the cells were adherent to the plastic surface and stained positive for CD14 and other markers specific for macrophages.

TCGA analyses. Data were obtained from the TCGA acute myeloid leukaemia database (version: 16 August 2016). The patients were classified into AML subtypes (FAB classification) M0 (undifferentiated acute myeloblastic leukaemia) ($n = 16$), M1 ($n = 42$), M2 ($n = 39$), M3 ($n = 16$), M4 ($n = 35$), M5 ($n = 18$), M6 ($n = 2$), M7 ($n = 3$); two cases were not classified by subtype. The mRNA levels of indicated genes were determined by RNA-seq (polyA+ IlluminaHiSeq). RESM-normalized counts are reported, and data were analysed and visualized with UCSC Xena (https://xena.ucsc.edu/). For analysis of overall survival, 160 patients with available survival data were separated into three groups based on whether they had high, moderate or low gene expression and then analysed by Xena Kaplan–Meier plot (http://xena.ucsc.edu/survival-plots/).

Flow cytometry. Primary antibodies including anti-human CD45-PE (BD Pharmingen, HI30, 1:100), CD45-FITC (BD Pharmingen, HI30, 1:100), CD45-APC (BD Pharmingen, HI30, 1:100), anti-human CD34-FITC (BD Pharmingen, 55582, 1:100), anti-human CD19-PE (eBioscience, HIB19, 1:100), anti-human CD20-PE (BD Pharmingen, 555623, 1:100), anti-human CD11b-APC (eBioscience, ICRF44, 1:100), anti-human LILRB4-APC (eBioscience, ZM4.1, 1:100), anti-human LILRB4-PE (Biolegend, ZM4.1, 1:100), anti-human CD14-APC (eBioscience, 61D3, 1:100), anti-human CD33-APC (Biolegend, P67.6, 1:100), anti-human CD4-APC (eBioscience, RPA-T4, 1:100), anti-human CD3-FITC (BioLegend, HIT3a, 1:100), anti-human CD3-Pacific blue (BD Pharmingen, SP34-2, 1:100) anti-human CD8-PE (BD Pharmingen, 555367, 1:100), anti-human CD28-APC (eBioscience, CD28.2, 1:100), anti-human CD40L-APC (eBioscience, 24-31, 1:100), anti-human PD1-APC (Biolegend, EH12.2H7, 1:100), anti-human TIM3-APC (eBioscience, F38-2E2, 1:100), anti-human TIGIT-APC (eBioscience, MBSA43, 1:100), anti-human LAG3-APC (eBioscience, 3DS223H, 1:100), anti-human FasL-PE (eBioscience, 24-31, 1:100), anti-uPAR-APC (Biolegend, VIM5, 1:100), anti-mouse CD3-APC (BioLegend, 17A2, 1:200), anti-mouse CD8a-PE (BioLegend, 53-6.7, 1:200), anti-mouse CD45-PE (BD Pharmingen, 30-F11, 1:200), anti-mouse CD49b-APC (eBioscience, DX5, 1:200), anti-mouse CD49f-PE (eBioscience, GoH3, 1:200), anti-mouse CD11b-APC (BioLegend, M1/71, 1:200), anti-mouse CD11b-PE (BioLegend, M1/71, 1:200), anti-mouse CD11c-APC (eBioscience, N418, 1:200), anti-mouse F4/80-APC (BioLegend, BM8, 1:200), anti-His-tag-APC (R&D Systems, AD1.1.10, 1:400), and IgG isotype-control-APC (eBioscience, P3.6.2.8.1, 1:400) antibodies were used. Cells were run on either Calibur for analysis or FACSAria for analysis and sorting. Flow data were analysed by Flowjo software. For analysis of human haematopoietic engraftment in NSG mice, a previously published protocol was followed[33,35,37]. Propidium iodide (PI) staining was used to exclude dead cells in analysis and sorting. For intracellular staining, we followed the two-step protocol for fixation/methanol from eBioscience. In brief, human primary AML cells were stained for surface expression of LILRB4 (anti-LILRB4-Alexa Fluor 647, Biolegend, ZM4.1, 1:100) and CD33 (anti-human CD33-FITC, Biolegend, HIM3-4, 1:100) and fixable cell viability dye eFluor 450 (Bioscience, Cat#65-0863-14, 1:100) followed by fixation (IC fixation buffer, eBioscience, Cat#00-8222) and methanol treatment. After that, cells were stained for intracellular antigens by anti-p-SHP-2 (Y580)-PE (Cell Signaling, Cat#13328S, 1:100), anti-pIKKα/β (S176/180) (16A6) (Cell Signaling, Cat#2697, 1:100), anti-NFκB (S529)-PE (eBioscience, B33B4WP, 1:100), anti-uPAR-PE (Biolegend, VIM5, 1:100), anti-arginase-1 (D4E3M) (Cell Signaling, Cat#93668, 1:100), rabbit IgG isotype control-PE (Cell Signaling, Cat#5742, 1:100), mouse IgG isotype control-PE (eBioscience, m2a-15F8, 1:100) and anti-rabbit IgG-PE (Jackson Immunoresearch Laboratory, Cat#111-116-144, 1:400) for flow cytometry analysis.

Virus construction and infection. For retrovirus packaging, plasmid constructs XZ201-IRES-GFP and XZ201-human *LILRB4* (h*LILRB4*)-IRES-GFP were mixed with PCL-ECO (2:1), followed by transfection into 293T cells using Lipofectamine 2000 (Invitrogen). For lentivirus packaging, CRISPR–Cas-9-based guide RNA (gRNA) constructs and other constructs for gene overexpression, including pLentiLox3.7-luciferase-IRES-GFP, ZsGreen-h*LILRB4* and ZsGreen-h*LILRB4*-intΔ, pLVX-*PLAUR*-IRES-tdTomato and pLVX-*ARG1*-IRES-tdTomato were mixed with psPAX2 and pMD2.G (Addgene) at a ratio of 4:3:1 and transfected into 293T cells using Lipofectamine 2000 (Invitrogen). Virus-containing supernatant was collected 48–72 h post-transfection and used for infection as previously described[38].

CRISPR–Cas9-based gene knockout in AML cells. Human AML cells were infected with doxycycline-inducible Cas9-expressing lentivirus (pCW-Cas9, Addgene 50661). After 1 μg/ml puromycin selection, surviving cells were infected with sgRNA-expressing lentivirus, produced by the plasmid modified from pSLQ1651 (Addgene 51024) by replacing the puro-mcherry with GFP for sorting. Scramble control sgRNA (sgRNA 5′-GAACGACTAGTTAGGCGTGTA-3′), *LILRB4* targeting sgRNA (sgRNA1 5′-TGTTACTATCGCAGCCCTGT-3′; sgRNA2 5′-GTAGGTCCCCCCGTGCACTG-3′; sgRNA3 5′-CCTGTGACCTCAGTGCACGG-3′), *APOE* targeting sgRNA (sgRNA1 5′-CTT TTGGGATTACCTGCGC-3′; sgRNA2 5′-AACTGGCACTGGGTCGCTTT-3′), *SHP1* targeting sgRNA (sgRNA1 5′-TAAGACCTACATCGCCAGCC-3′; sgRNA2 5′-GAAGAACTTGCACCAGCGTC-3′), *SHP2* targeting sgRNA (sgRNA1 5′-GAGACTTCACACTTTCCGTT-3′; sgRNA2 5′-TACAGTACTACAACTCAAGC-3′), *SHIP* targeting sgRNA (sgRNA1 5′-CACGCAGAGCGCGTATGCCC-3′; sgRNA2 5′-TGGCAACATCACCCGCTCCA-3′) which were designed by an online tool (http://crispr.mit.edu), were cloned into the sgRNA plasmid, individually. After being treated with 1 μg/ml doxycycline (Sigma, Cat#PHR1789) for 1 week, these cells were stained with anti-LILRB4 antibody and the LILRB4 negative cells were sorted as *LILRB4*^{KO} cells. For *APOE*^{KO}, *SHP1*^{KO}, *SHP2*^{KO} and *SHIP*^{KO} cells, GFP^+ cells were sorted into a 96-well plate as a single cell per well. After cell expansion, knockout cells were verified by western blotting. For in vivo induction of CRISPR–Cas9 to achieve gene knockout, we fed mice with doxycycline as described[39]. In brief, 7 days after Cas9/*LILRB4*-sgRNA-transfected THP-1 cell implantation, mice were treated with 2 mg per mouse of doxycycline via gavage daily for 5 days to achieve

Cas9 expression in engrafted leukaemia cells. The knockout was validated by flow cytometry.

Leukaemia cell and T cell co-culture assay. In the co-culture assay, human T cells (5×10^4 per well) isolated from health donor peripheral blood (PB009-1-0, Allcells) were mixed with irradiated (28 Gy) indicated human leukaemia cells in a U-bottom 96-well plate. For non-contact co-culture of T cells with leukaemia cells, leukaemia cells were cultured in the upper chamber of transwell inserts (pore size, 3 μM, #09-761-80, Thermo Fisher) in U-bottom 96 well-plates. T cells isolated from healthy donors were placed in the lower chambers of a 96-well transwell plate. Irradiated indicated leukaemia cells (E:T ratio = 2:1 if not indicated) were added to the upper chambers and treated with indicated antibodies, proteins and reagents. After culture with anti-CD3/CD28-coated beads (11161D, Thermo Fisher) and 50 U/ml rhIL-2 for 5–7 days, representative cells were photographed using an inverted microscope, and T cells were stained with anti-CD3 antibodies and analysed by flow cytometry.

For primary AML or B-ALL samples, patient leukaemia cells were sorted as $CD33^+$ and $CD19^+$ for AML and B-ALL, respectively. These leukaemia cells were cultured with autologous $CD3^+$ T cells from the same patient or allogeneic T cells from a health donor (E:T ratio = 2:1). After culture with anti-CD3/CD28-coated beads (11161D, Thermo Fisher) and 50 U/ml rhIL-2 for 14 days, representative cells were photographed using an inverted microscope, and T cells were stained with anti-CD3, anti-CD4 and anti-CD8 antibodies and analysed by flow cytometry.

For cytotoxicity assay, human $CD8^+$ T cells (5×10^4 per well) isolated from peripheral blood mononuclear cells (PBMCs) from a healthy donor were stimulated with anti-CD3/CD28/CD137-coated beads (11163D, Thermo Fisher) for 2 days in a 96-well plate. Then, indicated 5×10^3 leukaemia cells and 50 to 500 μg/ml anti-LILRB4 antibodies or control IgG were added. Cell numbers were determined on day 7 in triplicate wells. Alternatively, indicated leukaemia cells in indicated E:T ratios were cultured with T cells for 4–6 h in triplicate wells. Anti-CD3 and anti-CD8 were used to detect human CTL cells; indicated live THP-1 cells were positive for GFP and negative for PI. Cell supernatants from co-cultures of stimulated CTL cells and THP-1 cells treated with anti-LILRB4 or IgG were used to examine cytokine production using human cytokine arrays (AAH-CYT-6, RayBiotech).

For co-culture of mouse leukaemia and T cells, spleen cells from wild-type C57BL/6 mice were co-cultured with 2.5×10^4 irradiated (28 Gy) mouse leukaemia C1498 cells in a U-bottom 96 well-plate for 60 h. Anti-CD3/CD28-coated beads (11452D, Thermo Fisher), 50 U/ml recombinant human IL-2, and 5% serum from wild-type C57BL/6 mice or from *Apoe*^{KO} mice were added to the medium. In some experiments, 50 μg/ml lipid-bound APOE proteins (APOE–POPC) were added to the medium. The lipidation of APOE recombinant protein was conducted as described[40].

Transendothelial migration assays. To measure the ability of AML cells to migrate through endothelial cells, 3×10^5 HUVECs were cultured on a transwell membrane (pore size 8 μm). After 3 days, 1×10^5 indicated leukaemia cells were seeded in the upper chamber. In indicated experiments, leukaemia cells were treated with antibodies or proteins in the upper chamber. After 18 h, cells in the lower chamber were counted.

Short-term infiltration assay of leukaemia cells and homing assay of haematopoietic stem/progenitor cells. Cells (5×10^6 cells per mouse) were injected intravenously into NSG mice. Animals were treated with 10 mg/kg of anti-LILRB4 antibodies or control IgG immediately after injection of leukaemia cells. Mice were euthanized by CO_2 asphyxiation and death was assured by cervical dislocation after 20 h. Peripheral blood, bone marrow, liver and spleen were collected, and single-cell suspensions were examined by flow cytometry. CFSE, GFP or indicated markers such as anti-human CD45 and anti-human CD33 were used to detect target leukaemia cells in indicated experiments. Numbers of leukaemia cells in recipient liver, spleen and bone marrow are reported as a ratio relative to cell numbers in peripheral blood.

To test the infiltration ability of mouse leukaemia cells, 5×10^6 C1498 GFP h*LILRB4* cells or C1498 GFP cells were injected intravenously into wild-type C57BL/6 or *Apoe*-null mice. Mice were euthanized after 20 h. GFP was used to detect leukaemia cells by flow cytometry. The number of leukaemia cells in recipient liver, spleen and bone marrow were normalized to numbers in peripheral blood, and are reported as a ratio.

To test the homing ability of haematopoietic stem/progenitor cells (HSPCs), 1×10^7 human cord blood mononuclear cells were injected intravenously into an NSG mouse. Mice were treated with 10 mg/kg of anti-LILRB4 antibodies or control IgG immediately after injection of mononuclear cells and were killed after 20 h. Anti-human CD45 and anti-human CD34 were used to detect human HSPCs by flow cytometry. Similarly, to test the infiltration ability of normal human monocytes, 5×10^6 CD14-positive selected monocyte from health donor PBMC were labelled by CFSE and injected intravenously into an NSG mouse. Mice were treated with 10 mg/kg of anti-LILRB4 antibodies or control IgG immediately after injection of monocytes and were euthanized after 20 h. CFSE-positive cells were analysed by flow cytometry.

Innate immune cell depletion. NK cell depletion was done by intraperitoneal injection of 50 μl anti-asialo GM1 antibodies (CL8955, Cedarlane) 3 days before leukaemia cell implantation, which resulted in >90% depletion of $CD45^+CD49b^+$ NK cells in the circulation of NSG mice. Macrophages were depleted by treating NSG mice with clodronate (dichloromethylene bisphosphonate) liposomes (SKU8909, Clodrosome) (200 μl of stock solution 3 days before leukaemia cell implantation), resulting in >70% depletion of $CD45^+CD11b^+F4/80^+$ macrophages in the circulation of NSG mice. NSG mice were rendered neutropenic by intraperitoneal injection of 200μg anti-Ly-6G mAb (BP0075-1, Bioxcell) on days −3, −2, −1, and 0 after leukaemia cell implantation, resulting in >80% depletion of $CD45.1^+CD11b^+CD11c^-$ neutrophils in the circulation of NSG mice.

Human AML xenografts. Xenografts were performed essentially as described[2,3,6,7]. In brief, 6–8-week-old NSG mice were used for transplantation. Human leukaemia cells (1×10^6) were resuspended in 200 μl PBS for each mouse intravenous injection. Mice were immediately given 10 mg/kg of anti-LILRB4 antibodies or control IgG intravenously. Three to four weeks after transplantation, the peripheral blood, bone marrow, spleen and liver were assessed for engraftment. Leukaemia growth was monitored over time by luminescence imaging (maximum, 3×10^8 p/sec/cm^2/sr; min, 5×10^6 p/sec/cm^2/sr). For survival curve experiments, death was recorded when moribund animals were euthanized. For primary PDXs, each NSG mouse was given 5–10 × 10^6 human primary peripheral blood or bone marrow mononuclear cells, which contain leukaemia cells and other normal compartments such as normal haematopoietic stem progenitor cells and autologous T cells, via tail-vein injection. Mice were immediately given 10 mg/kg of anti-LILRB4 antibodies or control IgG intravenously and were treated twice a week until euthanization. For AML#11, mice were given 10 mg/kg of anti-LILRB4 antibodies or control IgG intravenously 7 days after leukaemia cell implantation and were treated twice a week until euthanization. Leukaemia growth was monitored over time by flow cytometry of human cells in peripheral blood. The presence of more than 1% of human leukaemia cells in mouse tissue was considered successful engraftment of primary AML cells. One to four months after transplantation, the peripheral blood, bone marrow, spleen, and liver were assessed for engraftment.

For the human PBMC (hPBMC)-humanized model, 1×10^7 hPBMCs were injected intravenously into each NSG mouse. Three weeks after implantation, mice had 30 to 50% engraftment of human T cells. At 3 weeks post implantation, 1×10^6 human AML THP-1 cells, including wild-type or *LILRB4*^{KO} THP-1 cells or THP-1 cells stably expressing luciferase (THP-1–Luc–GFP cells) were subcutaneously implanted. Mice were immediately given 10 mg/kg of anti-LILRB4 antibodies or control IgG intravenously and were treated twice a week until euthanization. Tumour growth was monitored over time by luminescence imaging (maximum, 1×10^8 p/sec/cm^2/sr; min, 5×10^6 p/sec/cm^2/sr). Tumour sizes were determined by caliper measure (width × width × length). For the inducible *LILRB4* knockout experiment, 1×10^6 Cas9/*LILRB4*-sgRNA-transfected THP-1 cells were injected intravenously into each NSG mouse, immediately followed by intravenous injection of 0.5×10^6 isolated human normal T cells from health donors. Seven days after THP-1 and T cell implantation, mice were treated with 2 mg per mouse of doxycycline via gavage daily for 5 days to achieve Cas9 expression in engrafted THP-1 cells. At 3 weeks post implantation, the peripheral blood, bone marrow, spleen, and liver were assessed for engraftment.

For the human cord blood (hCB) xenograft model, 2×10^4 $CD34^+$ hCB cells were injected intravenously into each NSG mouse. Six weeks after implantation, mice had 10 to 50% engraftment of human cells. THP-1 cells (1×10^6) that stably express luciferase were intravenously implanted. Mice were immediately given 10 mg/kg of anti-LILRB4 antibodies or control IgG intravenously. Tumour growth was monitored over time by luminescence imaging (maximum, 1×10^8 p/sec/cm^2/sr; min, 5×10^6 p/sec/cm^2/sr). Lineages of human normal blood cells were analysed by flow cytometry.

Mouse AML allograft. The procedure for mouse AML allografts was similar to that for human AML xenografts. In brief, 6–8-week-old wild-type C57BL/6 mice were used for transplantation. Mouse leukaemia cells (1×10^6) expressing human LILRB4 were resuspended in 200 μl PBS for each mouse for intravenous or subcutaneous implantation. Mice were given 10 mg/kg of anti-LILRB4-N297A antibodies or control IgG intravenously 7 days after leukaemia cell implantation and were treated twice a week until euthanization. Three weeks after transplantation, the peripheral blood, bone marrow, spleen and liver were assessed for engraftment. For subcutaneously implanted mice, tumour sizes were determined by caliper measure (width × width × length). For survival curve experiments, death was recorded when moribund animals were euthanized. For $CD8^+$ T depletion, 10 mg/kg anti-CD8 antibodies (YTS 169.4.2, Bioxcell) were injected intravenously 3 days after leukaemia cell implantation and mice were treated for an additional 2 times every 3 days. To determine whether anti-LILRB4 antibody treatment generated tumour-specific memory T cells against the tumour or against LILRB4,

we conducted adoptive transfer of spleen cells (5×10^6 cells per mouse) from anti-LILRB4 treated mice into normal recipient C57BL/6 mice. Four out of five transplanted mice rejected the control C1498-GFP mouse leukaemia cells, and these mice were not susceptible to rechallenge with threefold higher numbers (3×10^6 cells per mouse) of C1498-GFP leukaemia cells. Of 5 mice that received adoptive transfer of spleen cells from naive mice, none rejected the control C1498-GFP mouse leukaemia cells.

Chimeric receptor reporter assay. We constructed a stable chimeric receptor reporter cell system as described[3,4] to test the ability of a ligand to bind to the ECD of individual LILRBs, PirB, gp49B1 and LILRB4 site mutants and to trigger the activation or inhibition of the chimerically fused intracellular domain of paired immunoglobulin-like receptor β, which signals through the adaptor DAP-12 to activate the NFAT promoter. If an agonist or antagonist binds the ECD and activates or suppresses the chimeric signalling domain, an increase or decrease, respectively, in GFP expression is observed. A competition assay was used to screen LILRB4 blocking antibodies. In brief, APOE proteins (CI02, Novoprotein; 10 μg/ml) or human AB serum (10%, diluted in PBS) were pre-coated onto 96-well plates at 37 °C for 3 h. After two washes with PBS, 2×10^4 LILRB4 reporter cells were seeded in each well; meanwhile, indicated anti-LILRB4 antibodies were added into the culture medium. After 16 h, the percentage of GFP^+ reporter cells was analysed by flow cytometry. The threshold of activation is 2 times that of negative control treatment.

Fast protein liquid chromatography and mass spectrometry. Ten per cent human AB serum in PBS was loaded onto a 16/60 Superdex 200 gel filtration column and eluted with PBS and 2 mM EDTA. Eighty fractions (40 ml) were collected, and each fraction (0.5 ml) was analysed by chimeric receptor reporter assay. The active fractions (#26–30) were loaded onto PAGE-gel and processed to liquid chromatography with tandem mass spectrometry (LC–MS/MS) analysis (Orbitrap Elite) for protein identification in the UTSW proteomics core. Recombinant or purified proteins used for validation were ZA2G (MBS145455, MyBioSource), AMBP (13141-H08H1, Sino Biological), TTHY (12091-H08H, Sino Biological), PEDF (11104-H08H, Sino Biological), A2MG (MBS173010, MyBioSource), HEMO (MBS143111, MyBioSource), ANGT (MBS173525, MyBioSource), A1AT (MBS173006, MyBioSource), S100A9 (pro-814, Prospecbio), HORN (EBP08267, Biotrend USA), VTDB (CSB-EP009306HU, Biotrend USA), LRG1 (pro-141, Prospecbio), A1BG (RPE570Hu01, Cloud-Clone), CRSP3 (RD172262100, BioVendor), APOA1 (16-16-120101-LEL, Athens Research & Technology), APOA2 (16-16-120102, Athens Research & Technology), APOA4 (16-16-120104, Athens Research & Technology), APOB (16-16-120200, Athens Research & Technology), APOC1 (16-16-120301, Athens Research & Technology), APOC2 (16-16-120302, Athens Research & Technology), APOC3 (16-16-120303, Athens Research & Technology), hAPOE (16-16-120500, Athens Research & Technology), mAPOE (CJ05, Novoprotein), APOE2 (350-12, Peprotech), APOE3 (350-02, Peprotech), APOE4 (350-04, Peprotech), PODXL2 (1524-EG-050, R&D Systems), CD44 (12211-H08H, Sino Biological), HCK (PV6128, Thermo Fisher), VEGFR3 (10806-H08H, Sino Biological), NRG3 (16071-H08H, Sino Biological), PI16 (H00221476-P01, Novusbio), hMAG (8940-MG-050, R&D Systems), mMAG (8580-MG-100, R&D Systems), CNTF (303-CR-050, R&D Systems), ANGPTL-7 (914-AN-025/CF, R&D Systems), integrin-α1β1 (7064-AB-025, R&D Systems), integrin-α2β1 (5698-AB-050, R&D Systems), integrin-α2β3 (7148-AB-025, R&D Systems), integrin-α3β1 (2840-A3-050, R&D Systems), integrin-α4β1 (5668-A4-050, R&D Systems), integrin-α4β7 (5397-A3-050, R&D Systems), integrin-α5β1 (3230-A5-050, R&D Systems), integrin-α5β3 (3050-AV-050, R&D Systems), integrin-α5β5 (2528-AV-050, R&D Systems), integrin-α5β6 (CT039-H2508H, Sino Biological), mintegrin-α5β6 (CT051-M2508H, Sino Biological), integrin-α5β8 (4135-AV-050, R&D Systems), integrin-α6β4 (5497-A6-050, R&D Systems), integrin-α8β1 (CT016-H2508H, Sino Biological), integrin-α9β1 (5438-A9-050, R&D Systems), integrin-α10β1 (5895-AB-050, R&D Systems), integrin-α11β1 (6357-AB-050, R&D Systems), integrin-αEβ7 (5850-A3-050, R&D Systems), integrin-αXβ2 (CT017-H2508H, Sino Biological) and normal mouse serum (NS03L, Millipore sigma).

Bio-layer interferometry. Binding interaction analyses between LILRB4-Fc and APOE2, APOE3 or APOE4 were performed on the Octet RED96 (ForteBio, Pall). All interaction studies were performed with the protein A dip-and-read biosensors (ForteBio). All binding experiments were performed using the Octet Red and kinetics buffer at 30 °C. LILRB4-Fc coated biosensors (25 μg/ml LILRB4-Fc was loaded for 420 s) were washed in kinetics buffer before monitoring of association (300 s) and dissociation (600 s) of APOEs. Background wavelength shifts were measured from reference sensors that were loaded only with LILRB4-Fc.

Surface plasmon resonance. Biacore 2000 and CM5 chips were used to analyse binding of recombinant APOEs to the LILRB4 extracellular domain fused to hFc as described[2]. Recombinant protein A (Pierce) was pre-immobilized in two flow cells using the amine-coupling kit from GE. LILRB4-hFc was injected into one of the flow cells to be captured by protein A. Each binding sensorgram from the sample flow cell, containing a captured LILRB4-hFc, was corrected for the protein A-coupled cell control. Following each injection of an antigen solution, which induced the binding reaction, and the dissociation period during which the running buffer was infused, the protein A surface was regenerated by injection of regeneration solution containing 10 mM Na_3PO_4 (pH 2.5) and 500 mM NaCl. All captured LILRB4-hFc, with and without APOE bound, was completely removed, and another cycle begun. All measurements were performed at 25 °C with a flow rate of 30 μl/min.

Microscale thermophoresis. MST experiments were performed on a Monolith NT.115 system (NanoTemper Technologies) using 80% LED and 20% IR-laser power. Laser on and off times were set at 30 s and 5 s, respectively. Recombinant LILRB4–ECD protein (SinoBio) was labelled with 4488-NHS (NanoTemper Technologies) and applied at a final concentration of 5.9 nM. A twofold dilution series was prepared for unlabelled His–APOE (CI06, Novoprotein) in PBS, and each dilution point was similarly transferred to LILRB4–ECD solution. The final concentrations of His–APOE ranged from 0.36 nM to 12 μM. Samples were filled into standard-treated capillaries (NanoTemper Technologies) for measurement.

Western blotting and co-immunoprecipitation. Whole cells were lysed in Laemmli sample buffer (Sigma-Aldrich) supplemented with protease inhibitor cocktail (Roche Diagnostics). Samples were separated on SDS–PAGE gels (Bio-Rad) and transferred onto nitrocellulose membranes (Bio-Rad) for protein detection. Primary antibodies including anti-SHP-1 (Cell Signaling, 3759, 1:1,000), anti-phospho-SHP-1 Tyr564 (Cell Signaling, 8849, 1:500), anti-phospho-SHP-1 Tyr564 (Invitrogen, PA537708, 1:500), anti-SHP-2 (Cell Signaling, 3397, 1:1,000), anti-phospho-SHP-2 Tyr580 (Cell Signaling, 3703, 1:500), anti-SHIP1 (Cell Signaling, 2727, 1:1,000), anti-phospho-SHIP1 Tyr1020 (Cell Signaling, 3941, 1:500), anti-NFκB p65 (Cell Signaling, 8242, 1:1,000), anti-IKKα (Cell Signaling, 11930, 1:1,000), anti-IKKβ (Cell Signaling, 8943, 1:1,000), anti-phospho-IKKα/β Ser176/180 (Cell Signaling, 2697, 1:500), anti-IκBα (Cell Signaling, 4814, 1:1,000), anti-phospho-IκBα Ser32 (Cell Signaling, 2859, 1:500), anti-Lamin-B2 (Cell Signaling, 12255, 1:1,000), anti-arginase-1 (Cell Signaling, 9819, 1:1,000), anti-uPAR (Invitrogen, MON R-4-02, 1:500), anti-LILRB4 (Santa Cruz, sc-366213, 1:200), anti-APOE (Creative Diagnostics, DCABH-2367, 1:250), anti-β-actin (Sigma-Aldrich, A2066, 1:1,000) and anti-α-tubulin (Sigma-Aldrich, MABT205, 1:1,000), as well as horseradish peroxidase conjugated secondary antibodies (Cell Signaling, 7074, 1:1,000, and 7076, 1:1,000) and chemi-luminescent substrate (Invitrogen), were used. Specific cellular compartment fractionations were carried out using the NE-PER nuclear/cytoplasmic extraction kit (Thermo Fisher, 78833) or the plasma membrane protein extraction kit (Abcam, ab65400). Proteins from plasma membrane fraction were further incubated with anti-LILRB4 antibodies and dynabeads protein A (Thermo Fisher, 10001D) for further immunoprecipitation and western blotting.

Immunohistochemistry. Haematoxylin staining and immunostaining were performed on paraffin sections of tumours. Antibodies used were against LILRB4 (laboratory produced, 1:100), CD3 (Abcam, ab16669, 1:100), PD-1 (Thermo Fisher, J116, 14-9989-82, 1:100) and arginase-1 (Cell Signaling, 9819S, 1:100). The images were visualized using the Hamamatsu NanoZoomer 2.0-HT (Meyer instruments) and viewed in NPDview2 software (Hamamatsu).

Cytokine antibody array and arginase activity assay. To examine the secreted protein from leukaemia cells, conditioned media were applied to a human cytokine antibody array (AAH-CYT-1000, RayBio) for the semiquantitative detection of 120 human proteins. Image J (NIH) was used for quantification. Arginase activity was determined in condition medium of indicated leukaemia cells by a QuantiChrom Arginase assay kit (DARG100, BioAssay system).

RNA-seq analysis. RNA was purified from sorted cells with Qiagen RNeasy Mini Kit and then reverse-transcribed with SuperScript III Reverse Transcriptase (Invitrogen) according to the manufacturer's instructions. RNA-seq was performed at the UTSW Genomics and Microarray Core Facility. The cDNA was sonicated using a Covaris S2 ultrasonicator, and libraries were prepared with the KAPA High Throughput Library Preparation Kit. Samples were end-repaired, and the 3′ ends were adenylated and barcoded with multiplex adapters. PCR-amplified libraries were purified with AmpureXP beads and validated on the Agilent 2100 Bioanalyzer. Before being normalized and pooled, samples were quantified by Qubit (Invitrogen) and then run on an Illumina Hiseq 2500 instrument using PE100 SBS v3 reagents to generate 51-bp single-end reads. Before mapping, reads were trimmed to remove low-quality regions in the ends. Trimmed reads were mapped to the human genome (HM19) using TopHat v2.0.1227 with the UCSC iGenomes GTF file from Illumina.

Methods for data normalization and analysis are based on the use of 'internal standards' that characterize some aspects of the system's behaviour, such as technical variability, as presented elsewhere. Genes with $\log_2$ (fold change) > 2, $P < 0.01$ and RPKM >0.1 were deemed to be significantly differentially expressed between the two conditions and were used for pathway analysis and upstream transcription factor analysis. Pathway analysis was conducted using the DAVID

(https://david.ncifcrf.gov/tools.jsp). Upstream transcription-factor analysis was conducted using QIAGEN's Ingenuity tool (http://www.ingenuity.com/).

Molecular docking of LILRB4 with APOE. Docking of LILRB4 with APOE was performed on ZDOCKpro module of the Insight II package. The general protocol for running ZDOCK includes two consecutive steps of calculation described as geometry search and energy search, running in program ZDOCK and RDOCK, respectively. LILRB4 crystal structure (3P2T) and APOE3 structure (2L7B) were obtained from the Protein Data Bank database. The top 50 ZDOCK poses were submitted to RDOCK refinement. Poses with high scores in both ZDOCK and RDOCK were selected as candidate complex for LILRB4–APOE interaction analysis (Supplementary Table 2).

Statistical analyses. Representative data from four independent experiments or indicated independent samples are presented as dot plots (means ± s.e.m.) or as box-and-whisker plots (median values (line), 25th–75th percentiles (box outline) and minimum and maximum values (whiskers)). Statistical significance for two-sample comparisons was calculated by two-tailed Student's *t*-test. Statistical significance for survival was calculated by the log–rank test. The multivariate analysis of TCGA data was analysed by Cox regression. The difference was considered statistically significant if $P < 0.05$. NS, not significant; exact *P* values are shown. Pearson's correlation analyses were performed with RStudio software (the R Foundation).

Code availability. The custom code for Pearson's correlation analysis in RStudio is shown below.

```
setwd("~/file paths")
draw.graph = function(sam) {
file = paste(sam, ".tsv", sep = "")
df1 = read.table("file name A.tsv", sep = "\t", header = T)
df2 = read.table(file, sep = "\t", header = T)
df = merge(df1, df2, by.x = "sample", by.y = "sample")
df = df[!is.na(df[,2]),]
r = cor.test(df[,2], df[,3])
P = cor.test(df[,2], df[,3])$p.value
jpeg(filename = paste(sam,".jpeg", sep = ""),
width = 280, height = 280, units = "px", pointsize = 12,
quality = 75)
plot(df[[2]], df[[3]], xlab = "Title", ylab = sam,
pch = 16, cex = 1, col = "red",
main = paste("r = ",round(r$estimate,2), "\n","P = ",round(p,8), sep = ""))
reg = lm(df[[2]]~df[[3]])
abline(reg)
dev.off()
}
sample = c("file name B")
for (s in sample) {
draw.graph(s)
}
```

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

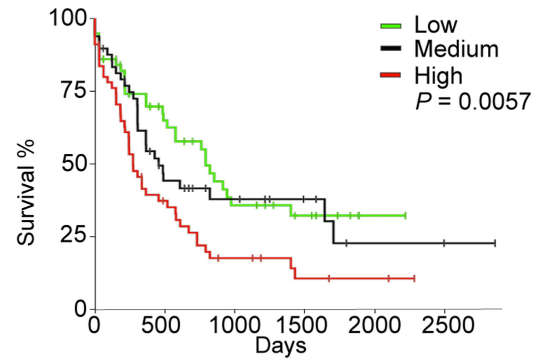
The TCGA datasets analysed are available in the UCSC Xena Browser (https://xena.ucsc.edu). The RNA-seq datasets generated in the current study have been deposited in NCBI SRA database with the SRA accession number SRP155049.

31. Piedrahita, J. A., Zhang, S. H., Hagaman, J. R., Oliver, P. M. & Maeda, N. Generation of mice carrying a mutant apolipoprotein E gene inactivated by gene targeting in embryonic stem cells. *Proc. Natl Acad. Sci. USA* **89**, 4471–4475 (1992).
32. Zheng, J. et al. Inhibitory receptors bind ANGPTLs and support blood stem cells and leukaemia development. *Nature* **485**, 656–660 (2012).
33. Kang, X. et al. The ITIM-containing receptor LAIR1 is essential for acute myeloid leukaemia development. *Nat. Cell Biol*. **17**, 665–677 (2015).
34. Deng, M. et al. A motif in LILRB2 critical for Angptl2 binding and activation. *Blood* **124**, 924–935 (2014).
35. Zheng, J. et al. Ex vivo expanded hematopoietic stem cells overcome the MHC barrier in allogeneic transplantation. *Cell Stem Cell* **9**, 119–130 (2011).
36. Lu, Z. et al. Fasting selectively blocks development of acute lymphoblastic leukemia via leptin-receptor upregulation. *Nat. Med*. **23**, 79–90 (2017).
37. Zhang, C. C., Kaba, M., Iizuka, S., Huynh, H. & Lodish, H. F. Angiopoietin-like 5 and IGFBP2 stimulate ex vivo expansion of human cord blood hematopoietic stem cells as assayed by NOD/SCID transplantation. *Blood* **111**, 3415–3423 (2008).
38. Zheng, J., Huynh, H., Umikawa, M., Silvany, R. & Zhang, C. C. Angiopoietin-like protein 3 supports the activity of hematopoietic stem cells in the bone marrow niche. *Blood* **117**, 470–479 (2011).
39. Cawthorne, C., Swindell, R., Stratford, I. J., Dive, C. & Welman, A. Comparison of doxycycline delivery methods for Tet-inducible gene expression in a subcutaneous xenograft model. *J. Biomol. Tech*. **18**, 120–123 (2007).
40. Denisov, I. G., Grinkova, Y. V., Lazarides, A. A. & Sligar, S. G. Directed self-assembly of monodisperse phospholipid bilayer Nanodiscs with controlled size. *J. Am. Chem. Soc*. **126**, 3477–3487 (2004).

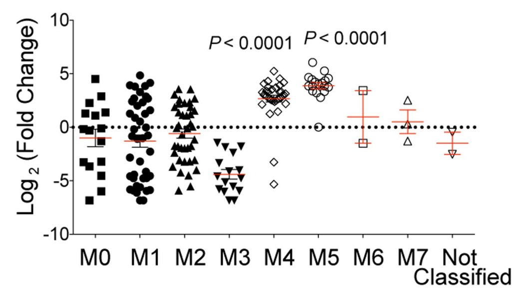
a

Gene Symbol	*P* Value
LILRB4	0.005735
LILRA5	0.01034
LILRB2	0.01183
PD1	0.01497
VISTA	0.02131
BTN1A1	0.02253
CRTAM	0.02889
LILRA6	0.04485
LILRB3	0.04821
PD-L2	0.05667
LILRA4	0.07504
CD96	0.08458
CD276	0.08495
CD28	0.09116
LILRA1	0.1202
BTN3A2	0.1617
LILRB1	0.1654
CD28H	0.2083
LAG-3	0.2223
LILRA2	0.2364
LAIR-1	0.2622
CD80	0.296
SLAMF2	0.2972
BTN3A1	0.3123
CD86	0.3344
BTNL3	0.3561
TIM-1	0.3561
SLAMF4	0.3864
PD-L1	0.3945
PVR	0.4403
PVRIG	0.4528
PDCD6	0.4679
ICOSLG	0.6398
BTN2A2	0.7636
BTN3A3	0.7739
CTLA-4	0.7806
BTLA	0.7943
LILRB5	0.7958
LAIR-2	0.7981
ICOS	0.8164
TIGIT	0.8254
CD226	0.8314
BTN2A1	0.8329
SLAMF5	0.8769
BTNL8	0.8938
TIM-3	0.8942
PVRL3	0.9493
PVRL2	0.956
VTCN1	Ø
BTNL2	Ø

b



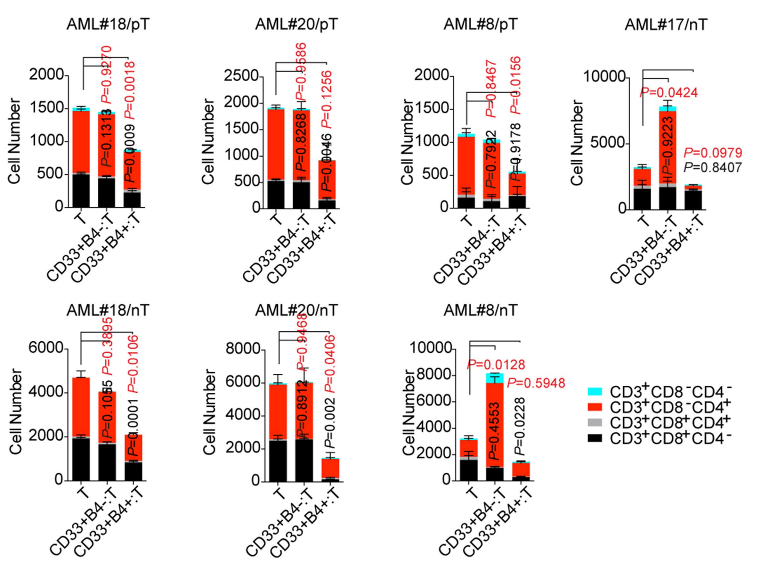
c



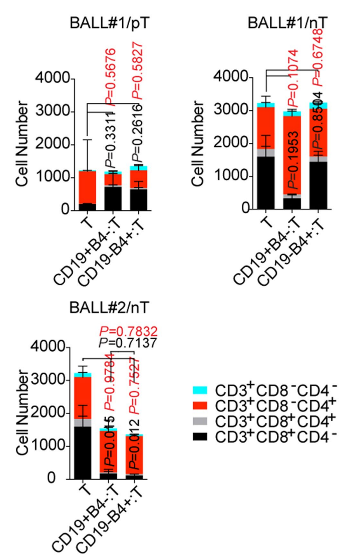
d

	Hazard Ratio (95% CI)	*P*-value
LILRB4 (medium vs. low)	1.518 (0.593, 3.887)	0.3843
LILRB4 (high vs. low)	2.339 (1.095, 4.999)	0.0283 *
Age (>=60 yrs vs. <60 yrs)	3.720 (1.895, 7.302)	0.0001 *
Cytogenetic (abnormal vs. normal)	0.761 (0.375,1.545)	0.4497
PML-RAR (mutated vs. normal)	0.835 (0.451, 1.548)	0.5675

e

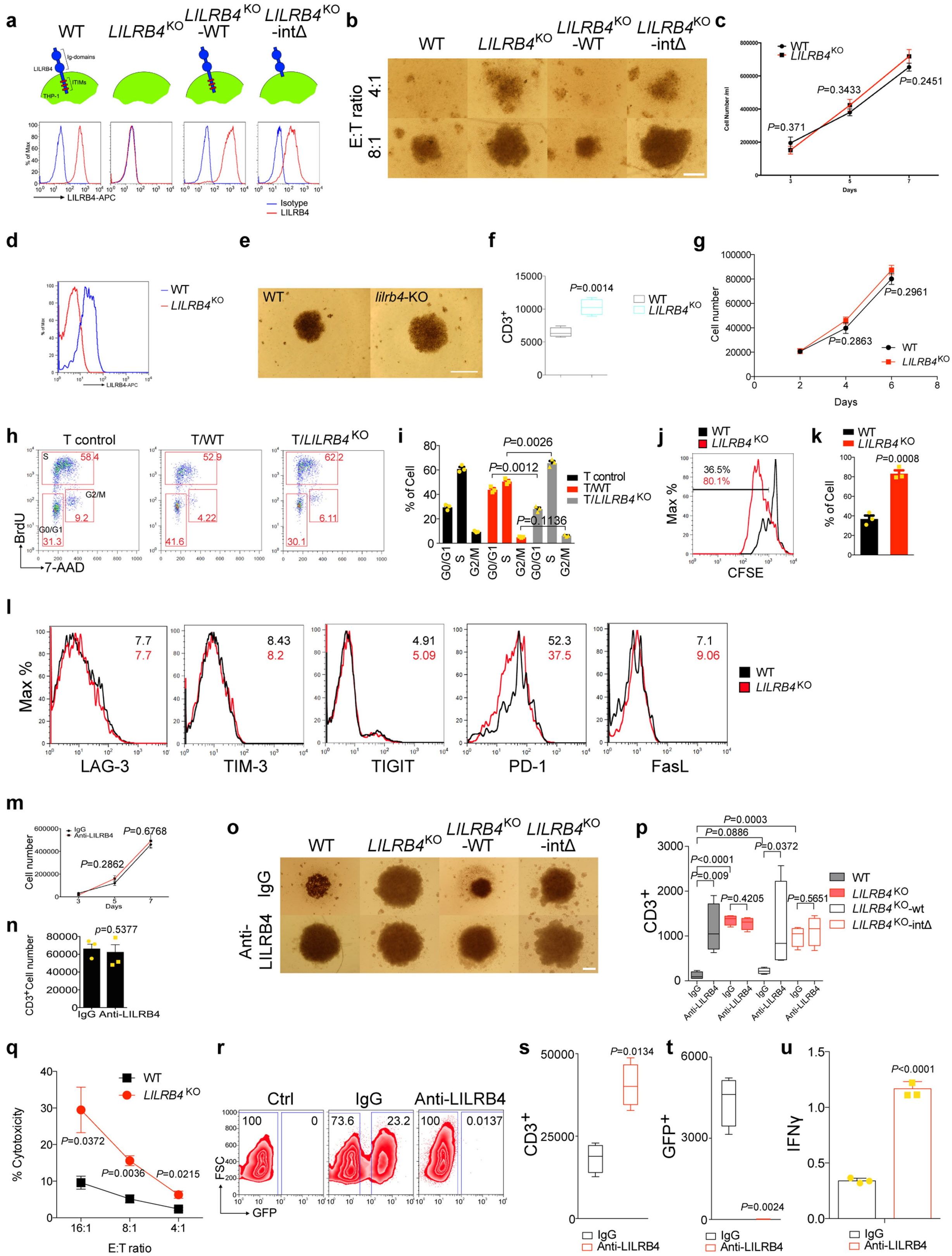


f



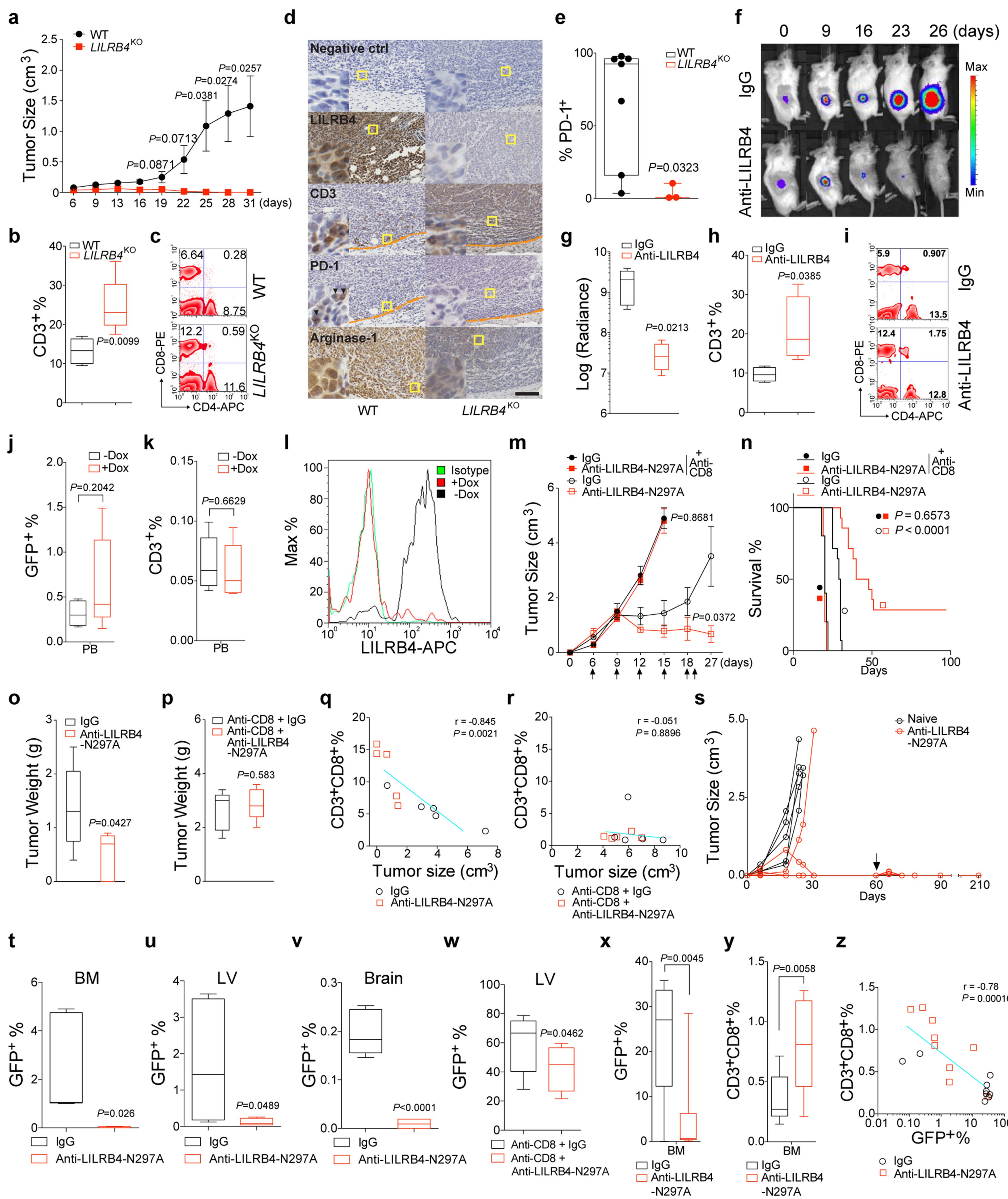
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | LILRB4 expression in patients with AML negatively correlated with overall survival and T cell proliferation. **a**, Analysis of correlation between mRNA levels of immune-modulating molecules and the overall survival of patients with AML ($n = 160$, divided into three groups based on gene expression) in TCGA database (https://xena.ucsc.edu) by Kaplan–Meier long-rank test. **b**, Kaplan–Meier analysis of correlations between *LILRB4* mRNA level and the overall survival of patients ($n = 160$) from the TCGA database, performed in Xena browser (https://xena.ucsc.edu). Low, $n = 57$; medium, $n = 48$; high, $n = 55$. The P value was from Kaplan–Meier log-rank test. **c**, mRNA expression data from the TCGA database were analysed as a function of the AML subtype of the patient. M0, $n = 16$; M1, $n = 42$; M2, $n = 39$; M3, $n = 16$; M4, $n = 35$; M5, $n = 18$; M6, $n = 2$; M7, $n = 3$; and two not-classified AML samples. Pairwise comparisons between M4 and each one of the other subtypes (all $P < 0.0001$), as well as between M5 and each one of the other subtypes (all $P < 0.0001$), using two-sample t-test. Mean and s.e.m. values are shown. **d**, A multivariable Cox regression analysis to assess the association, with adjustment for confounders that include age, cytogenetics and PML-RAR mutation in TCGA database. The total sample size was 79. $^*P < 0.05$ is considered significant. **e**, **f**, Autologous T cells isolated from individual patients with monocytic AML or B-ALL were incubated with irradiated *LILRB4*-positive or *LILRB4*-negative primary leukaemia cells from the same patients. pT, patient T cells. Allogeneic T cells isolated from healthy donors were incubated with irradiated *LILRB4*-positive or *LILRB4*-negative primary leukaemia cells from indicated patients with AML or B-ALL at an E:T of 10:1. nT, normal T cells. After culture with anti-CD3/CD28/CD137-coated beads and rhIL-2 for 14 days, T cells were stained with anti-CD3, anti-CD4, and anti-CD8 antibodies and analysed by flow cytometry. **e**, **f**, P values from two-tailed Student's t-test. P values in black indicate significance of $CD3^+CD8^+$ cells; P values in red indicate significance of $CD3^+CD4^+$ cells. $n = 2$ or 3 biologically independent samples with mean and s.e.m. See raw data for **e** and **f** in Source Data.



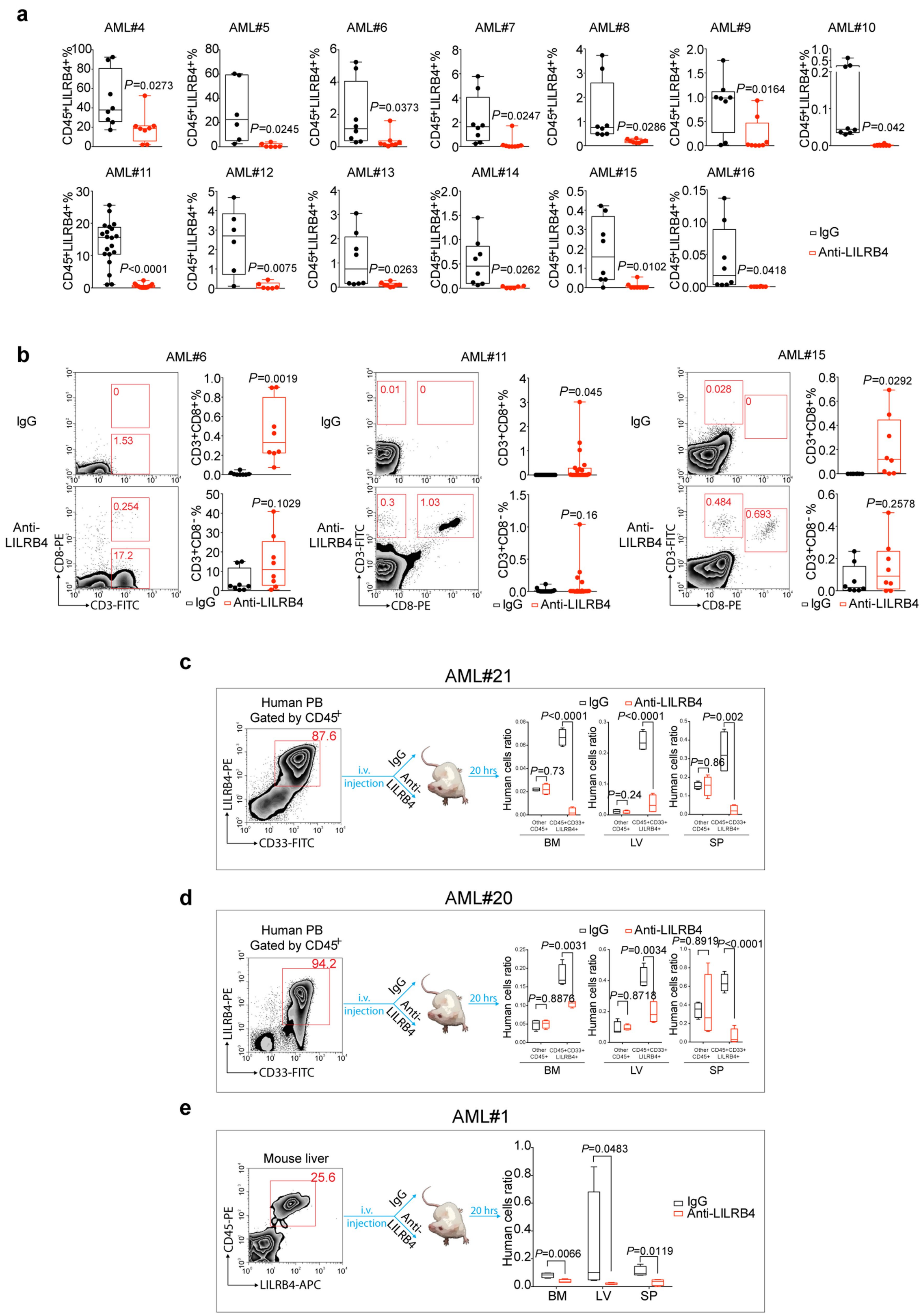
Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | LILRB4 suppresses T cell proliferation in vitro. **a**, Schematic of preparation of *LILRB4*-modulated THP-1 cells and examination of LILRB4 expression on the cell surfaces by flow cytometry. WT, THP-1 cells treated with scrambled control; *LILRB4*^{KO}, *LILRB4*-knockout THP-1 cells; *LILRB4*^{KO}-wt, forced expression of wild-type *LILRB4* in *LILRB4*^{KO} THP-1 cells; *LILRB4*^{KO}-intΔ, forced expression of intracellular domain-deleted mutant *LILRB4* in *LILRB4*^{KO} THP-1 cells. **b**, Loss of *LILRB4* on THP-1 cells reduces T cell suppression. Representative photograph of Fig. 1c (scale bar, 100 μm). **c**, Loss of *LILRB4* on THP-1 cells does not affect cell proliferation ($n = 3$ biologically independent samples with mean and s.e.m.). **d**, Examination of LILRB4 expression on cell surface of *LILRB4*^{KO} MV4-11 cells by flow cytometry. **e**, **f**, Loss of *LILRB4* on MV4-11 cells reduces T cell suppression. T cells isolated from healthy donors incubated in the lower chambers of a 96-well transwell plate with irradiated MV4-11 cells (E:T of 2:1) in the upper chamber separated by a membrane with 3-μm pores. After culture with anti-CD3/CD28-coated beads and rhIL-2 for 7 days, representative cells were photographed using an inverted microscope (scale bar, 100 μm) (**e**) and T cells were stained with anti-CD3 and analysed by flow cytometry (**f**). $n = 4$ biologically independent samples. **g**, Loss of *LILRB4* on MV4-11 cells does not affect cell proliferation ($n = 3$ biologically independent samples with mean and s.e.m.). **h**, **i**, T cells (E, effector cells) isolated from healthy donors were incubated with indicated irradiated THP-1 cells (T, target cells) without direct contact in transwells for 2 days. E:T = 2:1. T cells were treated with BrdU for 30 min followed by BrdU and 7-AAD staining for flow cytometry analysis. Representative flow cytometry plots are shown in **h** and the cell cycle status is summarized in **i**. T control, T cells cultured without THP-1 cells. $n = 3$ biologically independent samples with mean and s.e.m. **j**, **k**, T cells (E, effector cells) isolated from healthy donors were stained with CFSE and incubated with indicated irradiated THP-1 cells (T, target cells) without direct contact in transwells for 2 days. A representative flow cytometry plot is shown in **j** and the percentages of proliferating T cells indicated by CFSE-low staining is shown in **k**. $n = 3$ biologically independent samples with mean and s.e.m. **l**, LILRB4 increases PD-1 expression on T cells in coculture of leukaemia cells and T cells. T cells (E, effector cells) isolated from healthy donors were incubated with indicated irradiated THP-1 cells (T, target cells) in a non-contact manner for 5 days. E:T = 2:1. T cells were stained with anti-LAG-3, anti-TIM-3, anti-TIGIT, anti-PD-1 and anti-FasL antibodies for flow cytometry analysis. Representative flow cytometry plots and the mean of fluorescence intensities, at the right-upper corner (black, WT; red, KO), are shown. Experiments were performed three times with similar results. **m**, **n**, Anti-LILRB4 antibody had no effect on proliferation of THP-1 cells (**m**) or T cells (**n**). **m**, The growth of THP-1 cells during 7 days treatment with IgG or anti-LILRB4 antibody ($n = 3$ biologically independent samples with mean and s.e.m.). **n**, The numbers of human primary T cells after 5 days treatment with IgG or anti-LILRB4 antibody in vitro ($n = 3$ biologically independent samples with mean and s.e.m.). **o**, **p**, Primary T cells and irradiated THP-1 cells (E:T ratio, 2:1) were placed in the lower and upper chamber, respectively, and treated with 10 μg ml^{-1} control IgG or anti-LILRB4 antibodies. **o**, Representative photographs of T cells (scale bar, 100 μm). **p**, T cells stained with anti-CD3 and analysed by flow cytometry. $n = 4$ biologically independent samples. **q**, Primary T cells stimulated with anti-CD3/CD28/CD137-coated beads were co-cultured with wild-type or *LILRB4*^{KO} THP-1 cells with indicated E:T ratios for 4 h ($n = 3$ biologically independent samples with mean and s.e.m.). Cytotoxity of leukaemia cells was determined by PI staining in flow cytometry analysis. **r**–**u**, CD8^{+} T cells (5×10^4 cells) stimulated with anti-CD3/CD28/CD137-coated beads were co-cultured with 5×10^3 THP-1 cells that stably express GFP and treated with 100 μg ml^{-1} anti-LILRB4 antibodies or control IgG for 5 days. **s**, **t**, $n = 4$ biologically independent samples; **u**, $n = 3$ biologically independent samples with mean and s.e.m. Representative flow plots (**r**) of the percentages of T cells (GFP^{-}) and surviving leukaemia cells (GFP^{+}), and quantification of T cells (**s**), GFP^{+} leukaemia cells (**t**), and secretion of IFNγ (**u**), are shown. **b, d, e, h, j, o**, **r**, Experiments repeated independently three times with similar results. See Methods for definition of box plot elements in **f**, **p**, **s**, **t**. All *P* values were from two-tailed Student's *t*-test.



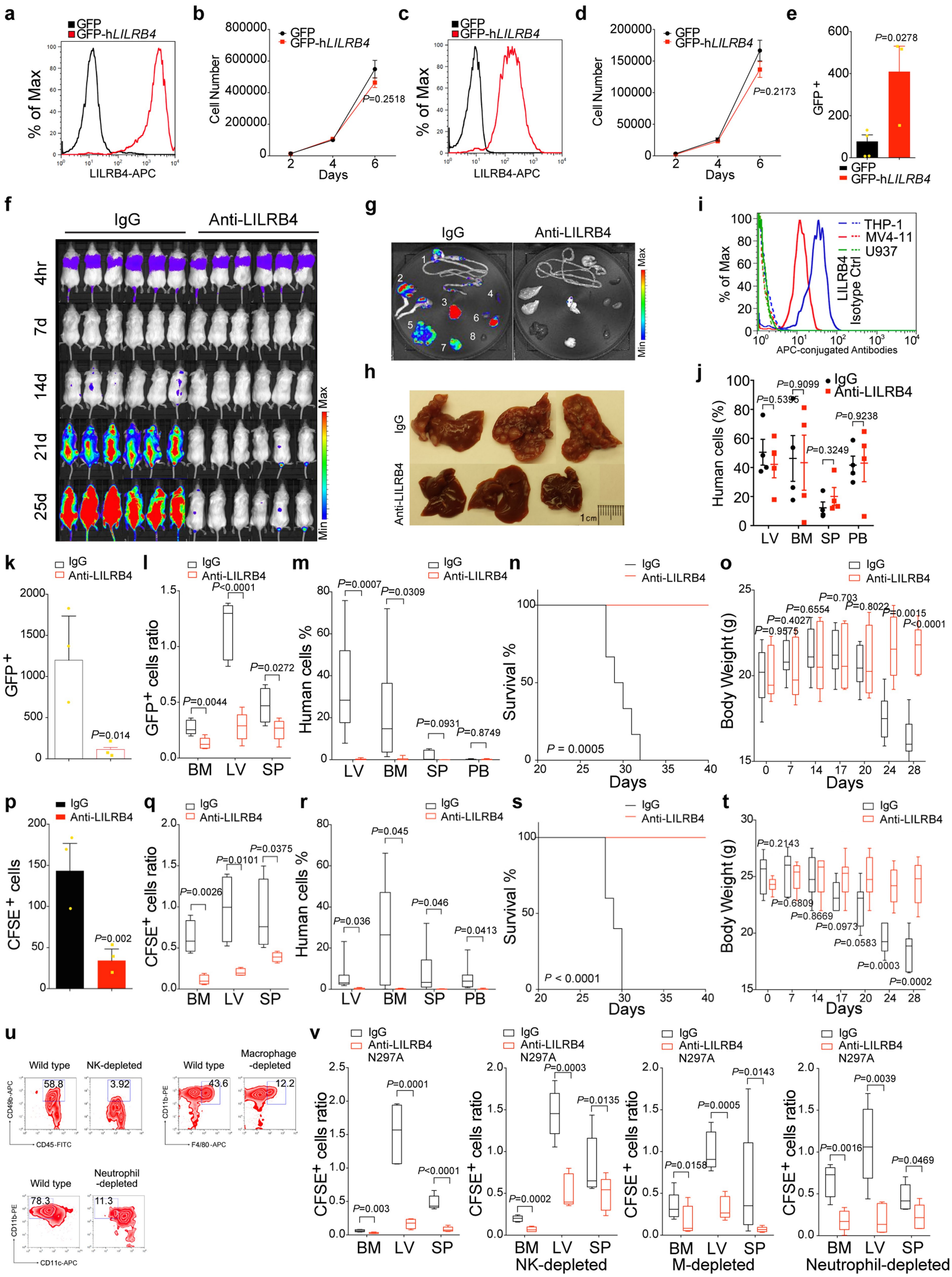
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Inhibition of LILRB4 reduces leukaemia development in humanized immunocompromised mice and syngeneic mice. **a**–**c**, Wild-type or *LILRB4*^{KO} THP-1 cells (3×10^6 cells per mouse) were subcutaneously implanted into hPBMC-repopulated NSG mice (WT, $n = 14$ mice with mean and s.e.m.; *LILRB4*^{KO}, $n = 10$ mice, mean and s.e.m.; also see Source Data). Tumour size (**a**), quantification of $CD3^+$ cells at day 31 in peripheral blood of recipient mice (**b**) and representative flow plots showing $CD4^+$ and $CD8^+$ T cells (**c**) are shown. **d**, **e**, LILRB4 increases PD-1 expression on tumour-infiltrated T cells. Wild-type or *LILRB4*^{KO} THP-1 cells were subcutaneously implanted into hPBMC-repopulated NSG mice. Three weeks after implantation, 7 out of 10 WT group mice had large tumours and 3 out of 10 knockout-group mice had tiny tumours. These tumours were dissected for immunohistochemistry and flow cytometry staining with anti-LILRB4, anti-CD3, anti-PD-1 and anti-ARG1 antibodies. Left corner images are magnified from yellow highlighted regions. In CD3 and PD-1 staining images, orange dashed lines indicate the tumour boundary. Black arrowheads indicate PD-1 positive cells. Scale bar, 100 μm. **e**, Tumours were dissected and cells in tumour region were stained with anti-CD3 and anti-PD-1 antibodies for flow cytometry analysis. The percentages of PD-1^+ T cells (ratio of PD-$1^+$$CD3^+$ cells to $CD3^+$ cells) were calculated. **f–i**, THP-1 cells were transplanted into hPBMC-repopulated NSG mice, and mice were treated with control IgG or anti-LILRB4 antibodies after 6 days (10 mg kg^{-1}; $n = 5$). Leukaemia development was monitored by luminescence imaging (**f**); luminescence flux (radiance) at day 26 (**g**; $n = 5$) and T cell numbers at day 26 in representative mice (**h**, **i**). **j**, **k**, Engraftment of human T cells and intravenously transplanted Dox-inducible *LILRB4*-knockout THP-1 cells (GFP^+) in NSG mice at day 7 before Dox administration ($n = 5$). **l**, Representative flow plot shows that LILRB4 was successfully deleted in engrafted leukaemia cells in bone marrow of Dox-fed mouse at the endpoint. n.s., not significant. **m**–**w**, Mouse AML C1498 cells (3×10^6 cells per mouse) that stably express LILRB4–IRES–GFP were subcutaneously implanted into C57BL/6 mice. Anti-LILRB4-N297A antibodies or control IgG were intravenously injected at 6, 9, 12, 15, 18 and 21 days after implantation of tumour cells. Two groups of mice were treated with anti-CD8 antibodies at 3, 6, 9 and 12 days after implantation of tumour cells to achieve $CD8^+$ T cell depletion. **m**, Tumour growth of subcutaneously implanted human LILRB4-expressing mouse AML C1498 cells (h*LILRB4* C1498) in C57BL/6 mice with anti-LILRB4-N297A antibodies or control antibody treatment ($n = 5$ mice). Also see Source Data. **n**, Survival curve of subcutaneous h*LILRB4* C1498 tumour-bearing mice ($n = 12$ mice). As for tumour size, anti-LILRB4 antibodies decreased the tumour weight (**o**, $n = 5$ mice) but did not do so in the absence of $CD8^+$ T cells (**p**, $n = 5$ mice). The percentage of $CD8^+$ T cells in spleen was significantly negatively correlated with tumour size (**q**, $n = 5$ mice) but not in the absence of $CD8^+$ T cells (**r**, $n = 5$ mice). **s**, Adoptive transplantation of spleen cells from control mice or tumour-bearing mice that were cured by anti-LILRB4-N297A treatment ($n = 5$ mice). Tumour size was monitored as a function of time. Arrow indicates day of rechallenge in mice that had eliminated leukaemia with three times the number of AML cells ($n = 4$ mice). Also see Source Data. Anti-LILRB4 antibodies reduced the infiltration of leukaemia cells into host tissues (**t**–**v**, $n = 5$ mice) and even $CD8^+$ cells were depleted (**w**, $n = 5$ mice). **x**–**z**, C57BL/6 mice were intravenously injected with human LILRB4-expressing mouse AML C1498 cells (3×10^6 cells per mouse) that expressed GFP. Anti-LILRB4-N297A antibodies ($n = 9$ mice) or control IgG ($n = 9$ mice) were intravenously injected at 6, 9, 12, 15 and 18 days after implantation of tumour cells. Anti-LILRB4 antibodies decreased the percentage of leukaemia cells in bone marrow (**x**). Anti-LILRB4 antibodies increased $CD8^+$ T cells (**y**). The percentage of $CD8^+$ T cells in bone marrow was significantly negatively correlated with the percentage of leukaemia cells (**z**). **c**, **i**, **l**, Experiments repeated independently three times with similar results. See Methods for definition of box plot elements in **b**, **e**, **g**, **h**, **j**, **k**, **o**, **p**, **t**–**y**. All *P* values (except **n**, long-rank test; and **q**, **r**, **z**, Pearson's correlation) from two-tailed Student's *t*-test.



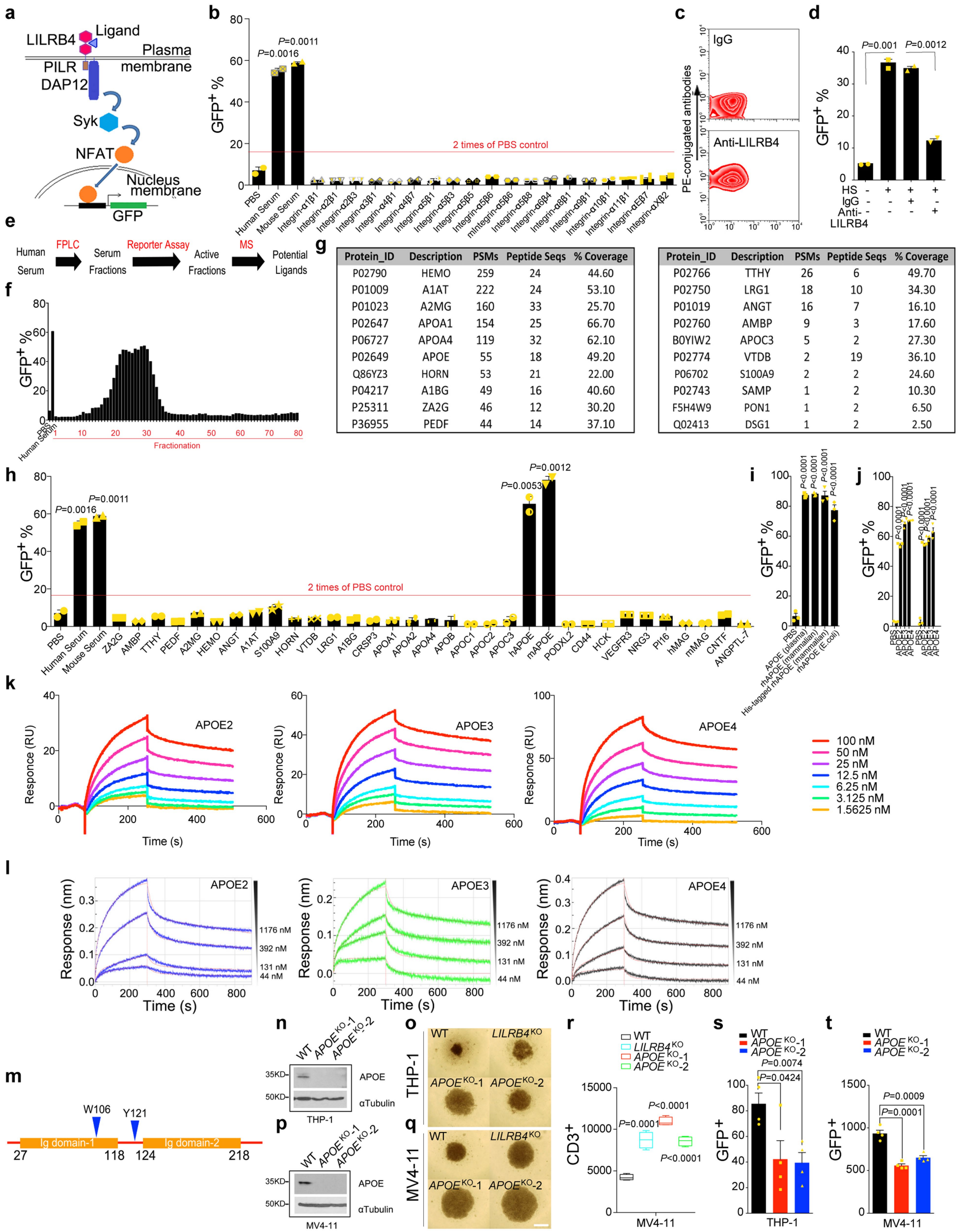
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Anti-LILRB4 antibodies reduce leukaemia development by restoring autologous T cells in PDX mice and inhibiting primary AML cell infiltration. a, Primary peripheral blood or bone marrow mononuclear AML cells (5×10^6 to 1×10^7 cells per mouse) from each of sixteen human patients (three shown in Fig. 1g–i, also see Supplementary Table 5) were injected into NSG mice followed by treatment with IgG or anti-LILRB4 antibodies (10 mg kg^{-1} twice a week by intravenous injection). Percentages of human $CD45^+LILRB4^+$ AML cells collected from haematopoietic tissues including bone marrow, spleen, liver and peripheral blood 2–4 months after transplantation, as determined by flow cytometry, are shown. **b**, Percentages of autologous human T cells collected from haematopoietic tissues including bone marrow, spleen, liver and peripheral blood 2–4 months after transplantation, as determined by flow cytometry; and representative flow plots of $CD3^+CD8^+$ T cells in bone marrow of mice in three PDXs. $n = 8$ biologically independent samples for all PDXs except AML#11 ($n = 20$ biologically independent samples) in **a**, **b**. **c**–**e**, Comparison of infiltration of human primary monocytic AML cells in NSG mice ($n = 5$ mice) after treatment with anti-LILRB4 antibody or IgG control. **c**, **d**, Primary human peripheral blood mononuclear cells from patients with monocytic AML were injected. The quantifications in **c** are also shown in Fig. 2l–n. **e**, Mouse liver cells with xenografted primary human monocytic AML cells (human $CD45^+LILRB4^+$ cells) were injected. See Methods for definition of box plot elements in **a**–**e**. All *P* values from two-tailed Student's *t*-test.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | LILRB4 promotes infiltration of AML cells. **a**, **c**, LILRB4 expression on mouse C1498 (**a**) or WEHI-3 (**c**) AML cells that stably express *Lilrb4* (also known as *Lilrb4a*). **b**, **d**, Forced expression of LILRB4 did not affect proliferation of mouse C1498 (**b**, $n = 3$ biologically independent samples with mean and s.e.m.) or WEHI-3 (**d**, $n = 3$ biologically independent samples with mean and s.e.m.) AML cells. **e**, Forced expression of human LILRB4 promoted transendothelial migration of mouse AML WEHI-3 cells ($n = 3$ biologically independent samples with mean and s.e.m.). **f**, NSG mice ($n = 6$ mice) were injected with 1×10^6 THP-1 cells followed immediately by IgG or anti-LILRB4 antibody treatment and were monitored by bioluminescence imaging. **g**, **h**, Anti-LILRB4 antibodies decreased AML cell infiltration into internal organs. Mice were killed at 21 days for ex vivo bioluminescence imaging of internal organs after transplantation of 1×10^6 luciferase-expressing THP-1 cells. Images of luminescence flux (radiance) from representative mice are shown (**g**). 1, GI tract; 2, legs; 3, lung; 4, spleen; 5, liver; 6, kidneys; 7, brain; 8, heart. Infiltrated leukaemia cells formed tumour nodules in liver (**h**). **i**, **j**, Anti-LILRB4 antibodies did not affect LILRB4-negative cancer cells. LILRB4 is expressed on THP-1 and MV4-11 human AML cells but not on U937 cells as analysed by flow cytometry (**i**). Isotype IgG was used as control. NSG mice were injected with U937 human AML cells, which do not express LILRB4, and then treated with anti-LILRB4 antibodies (**j**). IgG served as control antibodies. Mice were killed at day 25 post-transplant for analysis of LV, BM, SP and PB by flow cytometry. The presence of human AML cells was detected by anti-human CD45 antibody staining ($n = 4$ mice with mean and s.e.m.). **k**–**t**, Anti-LILRB4 antibodies decreased infiltration of THP-1 (**k–o**) or MV4-11 (**p–t**) human AML cells. Comparison of transendothelial migration abilities of GFP-expressing THP-1 (**k**) or CFSE-labelled MV4-11 (**p**) cells after treatment with anti-LILRB4 antibody or IgG control in a transwell assay ($n = 3$ biologically independent samples with mean and s.e.m.). Comparison of the homing abilities of GFP-expressing THP-1 or CFSE-labelled MV4-11 cells (5×10^6 per mouse) that were injected into NSG mice followed immediately by IgG or anti-LILRB4 antibody treatment. Numbers of leukaemia cells (GFP$^+$ in **l**, CFSE$^+$ in **q**) in LV, SP and BM normalized to that in PB as determined by flow cytometry 20 h after injection ($n = 5$ mice). NSG mice were injected with 1×10^6 THP-1 or MV4-11 cells followed immediately by IgG or anti-LILRB4 antibody treatment ($n = 6$ mice for THP-1 or 5 mice for MV4-11 xenografts). Percentages of MV4-11 cells (stained with anti-human CD45) as determined by flow cytometry in indicated organs at day 21 post-transplant (**m**, **r**), overall survival (**n**, **s**) and body weights as a function of time (**o**, **t**) are shown. **u**, Targeted immune cell populations were depleted in NSG mice. Representative flow cytometry plots demonstrating successful reduction of NK cell (CD45$^+$CD49b$^+$), macrophage (CD11b$^+$F4/80$^+$), and neutrophil (CD11b$^+$CD11c$^-$) frequency in NSG mice depleted of the respective immune cell subtype by treatment with anti-asialo GM1 antibodies, clodronate liposomes and anti-Ly6G antibodies, respectively, compared to non-depleted (wild-type) NSG mice. **v**, CFSE-labelled MV4-11 cells (5×10^6 per mouse) were injected into NSG mice in which the respective innate immune cells were depleted, followed immediately by IgG or anti-LILRB4-N297A antibody treatment ($n = 5$ mice). Numbers of leukaemia cells (CFSE positive) in LV, SP and BM normalized to that in PB at 20 h post-injection. **a**, **c**, **i**, **u**, Experiments repeated independently three times with similar results. See Methods for definition of box plot elements in **l**, **m**, **o**, **q**, **r**, **t**, **v**. All *P* values (except **n**, **s**, log-rank test) are from two-tailed Student's *t*-test.



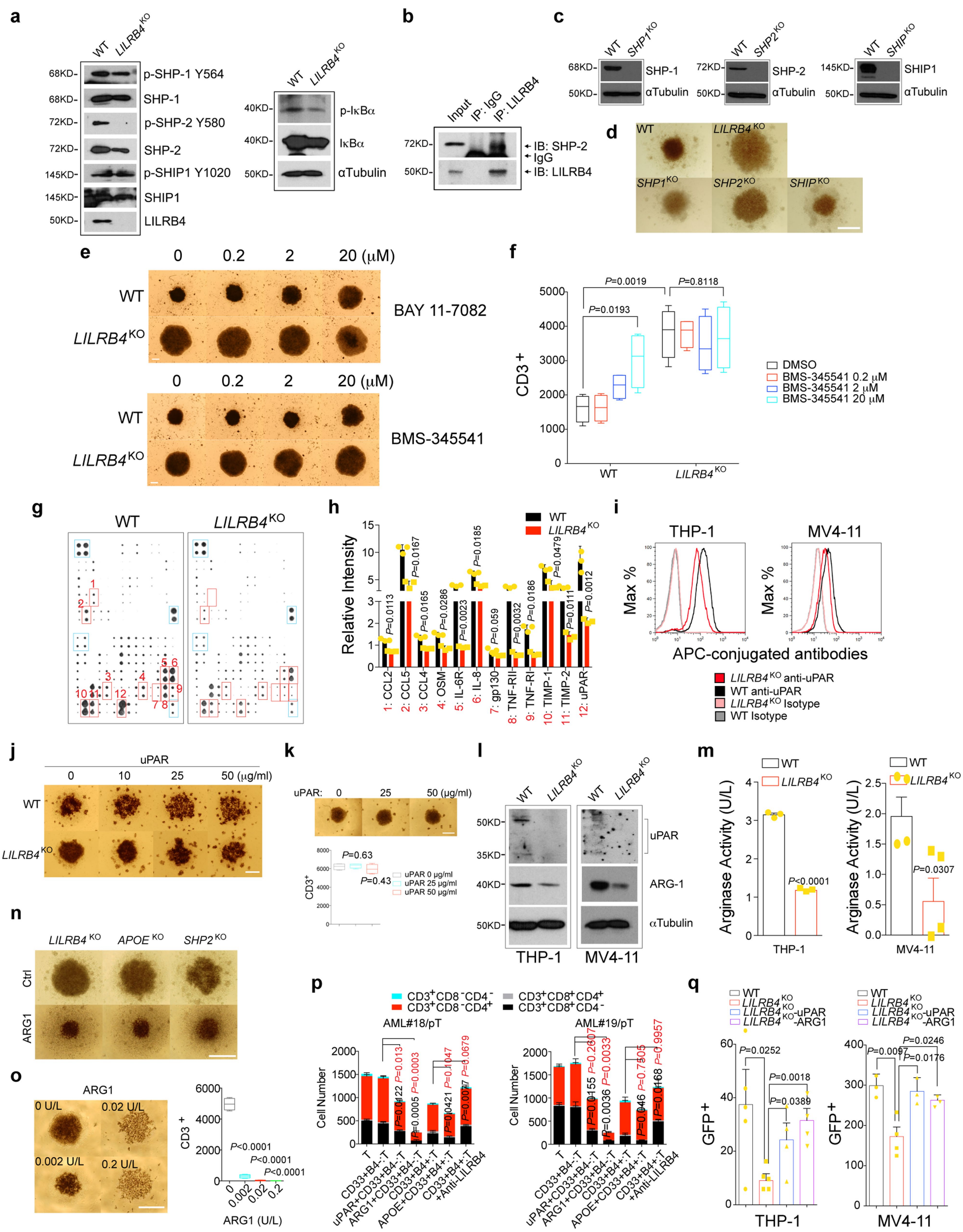
g

Protein_ID	Description	PSMs	Peptide Seqs	% Coverage
P02790	HEMO	259	24	44.60
P01009	A1AT	222	24	53.10
P01023	A2MG	160	33	25.70
P02647	APOA1	154	25	66.70
P06727	APOA4	119	32	62.10
P02649	APOE	55	18	49.20
Q86YZ3	HORN	53	21	22.00
P04217	A1BG	49	16	40.60
P25311	ZA2G	46	12	30.20
P36955	PEDF	44	14	37.10

Protein_ID	Description	PSMs	Peptide Seqs	% Coverage
P02766	TTHY	26	6	49.70
P02750	LRG1	18	10	34.30
P01019	ANGT	16	7	16.10
P02760	AMBP	9	3	17.60
B0YIW2	APOC3	5	2	27.30
P02774	VTDB	2	19	36.10
P06702	S100A9	2	2	24.60
P02743	SAMP	1	2	10.30
F5H4W9	PON1	1	2	6.50
Q02413	DSG1	1	2	2.50

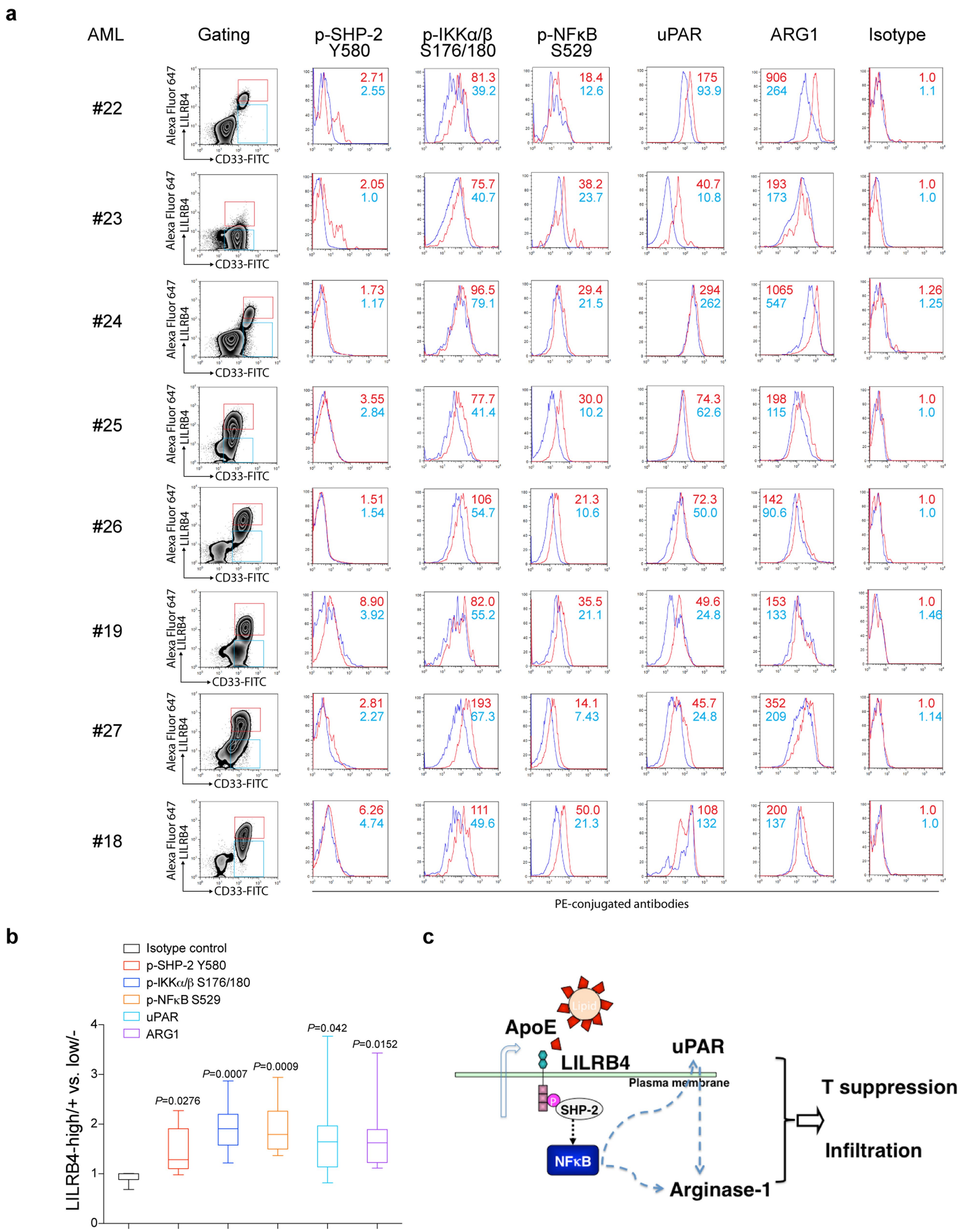
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | APOE induces LILRB4 activation to suppress T cells and support AML cell migration in vitro. a, Schematic of the LILRB4 reporter system. **b**, Human and mouse integrin heterodimer proteins cannot activate the LILRB4 reporter ($n = 3$ biologically independent samples with mean and s.e.m.). Human and mouse sera were used as positive controls. The threshold of activation is twice that of negative control treatment. **c**, Flow cytometry demonstrating that anti-LILRB4 antibody binds to human LILRB4 reporter cells. **d**, LILRB4 activation as indicated by percentage of GFP^+ cells in the presence and absence of 10% human serum (HS) with or without anti-LILRB4 antibody or control IgG ($n = 3$ biologically independent samples with mean and s.e.m.). **e**, Flow chart of ligand identification of potential ligands of LILRB4 in human serum. **f**, Fractionation of LILRB4-stimulating activities from human serum by fast protein liquid chromatography. The positive control was 10% human serum. **g**, Proteins identified from the LILRB4 stimulating fractions by mass spectrometry. PSMs, peptide spectrum matches. **h**, Both human and mouse APOE proteins can activate LILRB4 reporter ($n = 3$ biologically independent samples with mean and s.e.m.). Human and mouse sera were used as positive controls. The threshold of activation is twice that of negative control treatment. **i**, APOE proteins from different sources all activate LILRB4. APOE (20 μg ml^{-1}) purified from human plasma, His-tagged or tag-free recombinant human APOE (rhAPOE) (20 μg ml^{-1}) expressed by 293T mammalian cells, or rhAPOE (20 μg ml^{-1}) expressed by bacteria all activate the LILRB4 reporter. These APOE all represent human APOE3 ($n = 3$ biologically independent samples with mean and s.e.m.). **j**, APOE2, APOE3 and APOE4 all activate the LILRB4 reporter ($n = 3$ biologically independent samples with mean and s.e.m.). Forty micrograms per millilitre APOEs were coated on plates or directly added in cell culture medium (soluble). **k**, **l**, Three APOE isoforms bind to human LILRB4. **k**, Binding kinetics of APOE2, 3 and 4 to LILRB4-Fc were measured using SPR. LILRB4-Fc was immobilized on protein A biosensor tips and incubated with APOE concentrations ranging from 1.5625 nM to 100 nM. **l**, Binding kinetics of APOE2, 3, and 4 to LILRB4-Fc were measured using bio-layer interferometry (Octet). LILRB4-Fc was immobilized on protein A biosensor tips and incubated with APOE concentrations ranging from 44 nM to 1,176 nM. **m**, As shown in Fig. 3h, mutation of W106 and Y121, located in the first immunoglobulin domain and in the linker between two immunoglobulin domains, respectively, significantly reduced activation of LILRB4 by APOE. **n**, **p**, Examination of APOE expression in *APOE*-knockout THP-1 and MV4-11 cells by immunoblots. Primary T cells and irradiated THP-1 or MV4-11 cells (E:T = 2:1) were incubated in the lower and upper chambers, respectively. T cells were photographed (**o**, **q**, scale bar, 100 μm) and quantified by flow cytometry (Fig. 3i and **r**, $n = 4$ biologically independent samples) after 7 days. **s**, **t**, Loss of APOE suppresses transendothelial migration of human AML THP-1 and MV4-11 cells ($n = 4$ biologically independent samples with mean and s.e.m.). **c**, **k**, **l**, **n–q**, These experiments were repeated independently three times with similar results. See Methods for definition of box plot elements in **r**. All P values are from two-tailed Student's t-test.

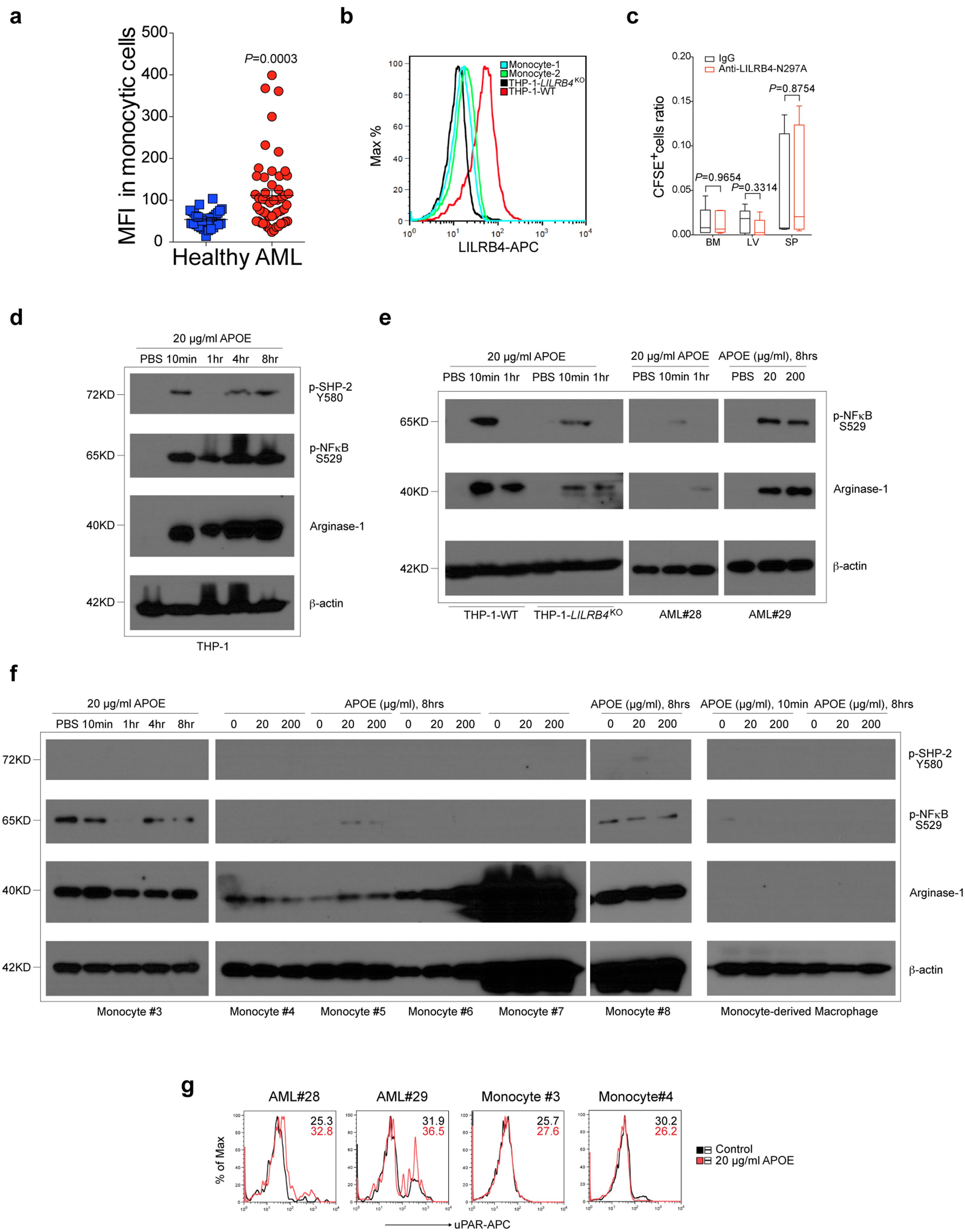


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | LILRB4 upregulates phosphorylation of SHP-2, NFκB signalling and expression of uPAR and ARG1 to suppress T cell activity and support leukaemia migration. **a**, Phosphorylated SHP-2 and IκBα were downregulated upon *LILRB4* knockout in MV4-11 cells. **b**, Co-immunoprecipitation demonstrated that LILRB4 interacts with SHP-2 in THP-1 cells. **c**, *SHP1*, *SHP2* and *SHIP* were individually knocked out by CRISP–Cas9 in THP-1 cells as detected by western blotting. **d**, Primary T cells and irradiated THP-1 cells (E:T = 2:1) were cultured in the lower and upper chambers, respectively. T cells were photographed (scale bar, 100 μm) after 7 days. **e**, **f**, Two different NFκB inhibitors restored T cell proliferation from the suppression by THP-1 cells in an LILRB4-dependent manner ($n = 4$ biologically independent samples). THP-1 cells were pretreated with various doses of NFκB inhibitors for 1 h. Primary T cells and irradiated pretreated THP-1 cells (E:T = 2:1) were cultured in the lower and upper chambers, respectively. T cells were photographed (**e**, scale bar, 100 μm) and analysed by flow cytometry (**f**) after 7 days. **g**, **h**, Loss of *LILRB4* decreased secreted protein production in THP-1 cells as determined by a human cytokine antibody array (**g**) and the blot intensities were quantified by ImageJ software (**h**, $n = 3$ biologically independent samples with mean and s.e.m.). Red boxes indicate proteins that were changed upon *LILRB4* knockout; blue boxes indicate positive controls. **i**, Surface uPAR was downregulated in *LILRB4*^{KO} THP-1 and MV4-11 AML cells. **j**, T cells were incubated with irradiated indicated THP-1 cells supplemented with indicated concentrations of recombinant uPAR proteins for 7 days. T cells were photographed. **k**, T cells isolated from healthy donors were cultured with anti-CD3/CD28-coated beads and rhIL-2 and supplemented with indicated concentrations of uPAR proteins for 3 days ($n = 4$ biologically independent samples). Representative cells were photographed using an inverted microscope and T cells were analysed by flow cytometry. **l**, Expression of uPAR and ARG1 is downregulated in in *LILRB4* knockout THP-1 and MV4-11 AML cells. **m**, Arginase activity as determined by a colorimetric method (DARG100, BioAssay system) was decreased in condition medium of *LILRB4*^{KO} THP-1 and MV4-11 cells ($n = 3$ biologically independent samples with mean and s.e.m.). **n**, Primary T cells and irradiated indicated THP-1 cells (E:T = 2:1) were incubated in the lower and upper chambers, respectively, and were supplemented with 0.002 U/l recombinant ARG1 proteins for 7 days. T cells were photographed. **o**, T cells isolated from healthy donors were cultured with anti-CD3/CD28-coated beads and rhIL-2 and supplemented with indicated concentrations of ARG1 proteins for 3 days ($n = 4$ biologically independent samples). Representative cells were photographed using an inverted microscope and T cells were analysed by flow cytometry. **p**, Autologous T cells isolated from individual patients with monocytic AML were incubated with irradiated *LILRB4*-positive or *LILRB4*-negative primary leukaemia cells from the same patients at an E:T of 10:1, supplemented with recombinant anti-LILRB4 antibodies, APOE-VLDL, uPAR or ARG1. pT, patient T cells. After culture with anti-CD3/CD28/CD137-coated beads and rhIL-2 for 14 days, T cells were stained with anti-CD3, anti-CD4, and anti-CD8 antibodies and analysed by flow cytometry. $n = 3$ biologically independent samples with mean and s.e.m. **q**, Supplementation of recombinant uPAR or ARG1 to the medium rescued the decrease in transmigration ability of *LILRB4*^{KO} THP-1 or *LILRB4*^{KO} MV4-11 cells across endothelium ($n = 3$ biologically independent samples with mean and s.e.m.). Scale bar, 100 μm. **a–e**, **g**, **i**, **j**, **l**, **n**, Experiments repeated independently three times with similar results. See Methods for definition of box plot elements in **f**, **k**, **o**. All P values from two-tailed Student's t-test. See raw data for **p** in Source Data.

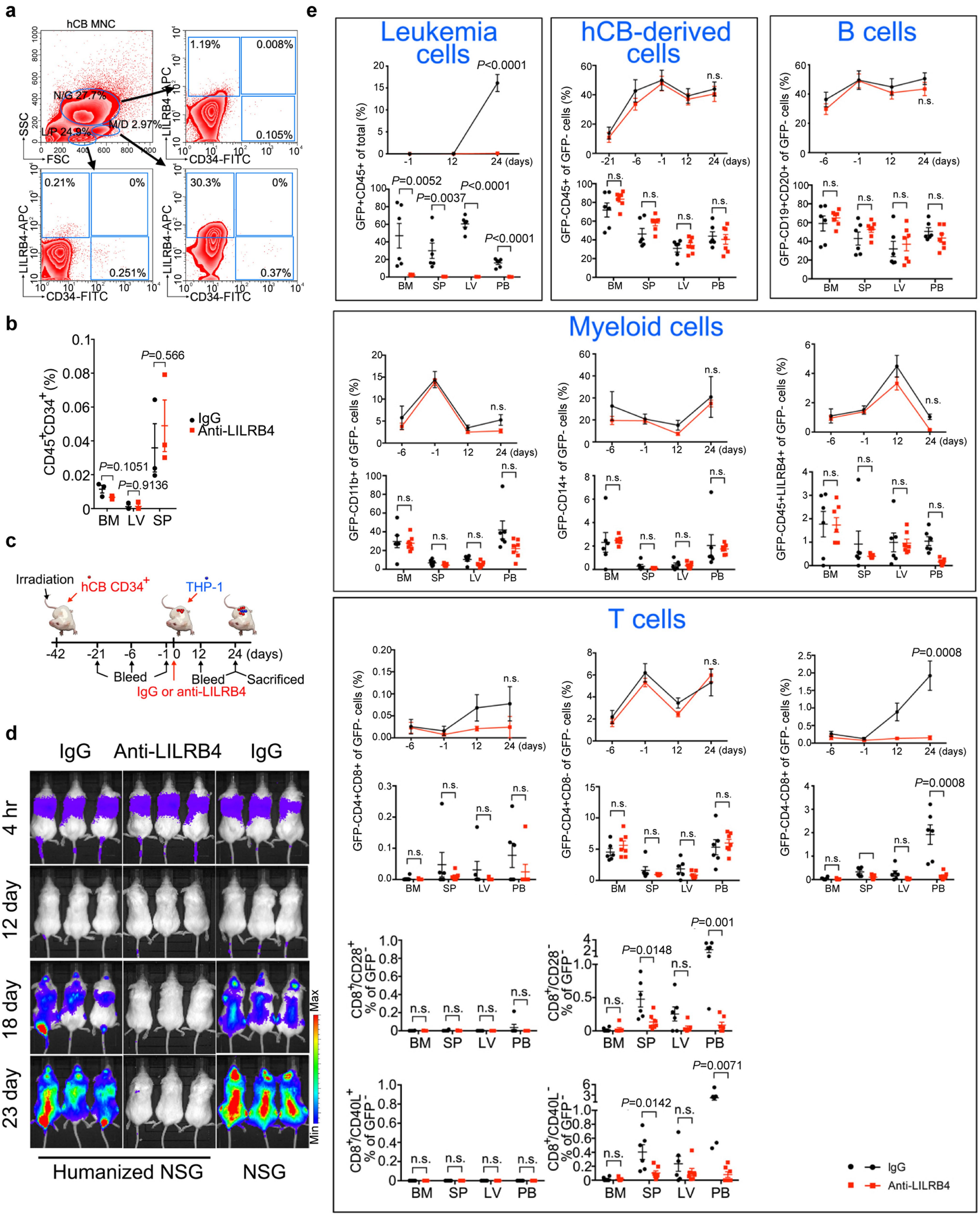


Extended Data Fig. 8 | Detection of SHP-2–NFκB signalling and uPAR and ARG1 expression in primary human monocytic AML cells. a, LILRB4-positive or -high $CD33^+$ AML cells (red box) and LILRB4-negative or -low $CD33^+$ AML cells (blue box) were gated for further intracellular staining of SHP-2 phosphorylated at Y580, IKKα/β phosphorylated at S176/S180, NFκB phosphorylated at S529, uPAR and ARG1. Isotype IgG was used as negative control. Red numbers indicate MFI of LILRB4-positive or -high $CD33^+$ AML cells; blue numbers indicate MFIs of LILRB4-negative or -low $CD33^+$ AML cells. This experiment was repeated with eight individual patient samples with similar results. **b**, Quantification of individual staining in LILRB4-positive or -high $CD33^+$ AML cells versus in LILRB4-negative or -low $CD33^+$ AML cells. $n = 8$ independent patients; see Methods for definition of box plot elements. P values from two-tailed Student's t-test. **c**, Schematic for the mechanisms by which LILRB4 suppresses T cells and promotes leukaemia infiltration.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Comparison of LILRB4-mediated intracellular signalling in leukaemia cells and in normal haematopoietic cells. **a**, Comparison of LILRB4 surface expression on normal monocytes from healthy donors ($n = 25$ individual donors with mean and s.e.m.) and neoplastic monocytic cells from patients with AML ($n = 53$ individual patients with mean and s.e.m.). **b**, Comparison of LILRB4 surface expression on normal monocytes from two healthy donors and on wild-type and *LILRB4*^{KO} THP-1 cells. This experiment was repeated independently three times with similar results. **c**, Anti-LILRB4 antibody did not affect homing ability of normal monocytes. Human normal monocytes (**b**) selected through CD14-positive selection. These isolated monocytes were pooled and stained by CFSE. After staining, monocytes (5×10^6 for each mouse) were injected into NSG mice followed immediately by antibody treatment, and then the mice ($n = 4$ mice, see Methods for definition of box plot elements) were killed 20 h after transplant. The number of CFSE$^+$ cells in liver, spleen and bone marrow were normalized to that in peripheral blood as determined by flow cytometry. **d**, **e**, APOE activates LILRB4 intracellular signalling in leukaemia cells. Indicated THP-1 cells and primary AML (M5) cells were serum-starved overnight and then treated with the indicated concentration of human recombinant APOE protein for the indicated time. Phospho-SHP-2, phospho-NFκB and ARG1 were examined by western blotting. **f**, The effect of APOE on normal monocytes or in vitro differentiated macrophages. Normal monocytes were isolated from healthy donors and macrophages were derived from these monocytes after one week of differentiation in vitro. Cells were serum-starved overnight and then treated with the indicated concentrations of human recombinant APOE protein for the indicated times. Phospho-SHP-2, phospho-NFκB and ARG1 were examined by western blotting. **g**, APOE induces uPAR upregulation on AML cells. Normal monocytes were isolated from healthy donors. Indicated primary AML cells and normal monocytes were serum-starved overnight and then treated with 20 $\mu g\ ml^{-1}$ human recombinant APOE protein for eight hours. Surface uPAR was examined by flow cytometry. Representative flow plots are shown and MFIs are shown in top right corner (black, PBS control; red, APOE treatment). Experiments were performed three times with similar results. *P* values from two-tailed Student's *t*-test.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Anti-LILRB4 does not affect engraftment of normal haematopoietic cells. a, LILRB4 and CD34 co-staining patterns for representative samples of human cord blood mononuclear cells (hCB MNCs). N/G, neutrophils and granulocytes; M/D, monocytes, macrophages and dendritic cells; L/P, lymphocytes, haematopoietic stem and progenitor cells. This experiment was repeated independently three times with similar results. **b**, Anti-LILRB4 antibody did not affect homing ability of normal haematopoietic progenitor cells. hCB MNCs (1×10^7) were injected into NSG mice followed immediately by antibody treatment, and then the mice ($n = 3$ mice with mean and s.e.m.) were killed 20 h after transplant. The number of $CD45^+CD34^+$ HSCs in liver, spleen, and bone marrow were normalized to that in peripheral blood as determined by flow cytometry. **c**–**e**, Anti-LILRB4 antibodies inhibited leukaemia development in hCB-humanized NSG mice. **c**, Schematic of the experiment to test whether anti-LILRB4 antibody inhibits leukaemia development in hCB-humanized NSG mice. **d**, Leukaemia development was monitored over time by luminescence imaging. This experiment was repeated independently twice with similar results. **e**, Frequency of engrafted leukaemia, normal human cells, including human B cells, human myeloid cells and human T cells in peripheral blood over time and haematopoietic tissues of hCB-humanized mice at 24 days after leukaemia transplantation. $n = 3$ mice with mean and s.e.m. All P values from two-tailed Student's t-test.

CAREERS

SOCIAL Follow us on Twitter at **twitter.com/naturejobs**

SHARE Tell us your career story at **naturecareerseditor@nature.com**

FAILURE Why it doesn't spell the end of your career **go.nature.com/2nmcbdv**

ILLUSTRATION BY SEÑOR SALME

JOB SATISFACTION

Paths less travelled

For many, a career in academia is the goal, but other options could prove more rewarding.

BY CHRIS WOOLSTON

Many science students and junior researchers continue to aspire to a career in academia, a dream that has persisted for generations. But *Nature*'s biennial survey of salary and job satisfaction in the global science community underscores an important reality: there is a vast number of career opportunities for scientists beyond academic research, and some of those options might be more rewarding, whether emotionally, financially or both.

Nature's survey — for which fieldwork was conducted between June and July 2018 by Shift Learning, a London-based research consultancy — drew responses from 6,413 self-selected readers from around the world. (Responses from people who hadn't gone beyond an undergraduate degree were filtered out, leaving a sample of 4,334.) Nearly 40% of respondents live in North America, 35% are in Europe and 16% are in Asia. *Nature* also heard from researchers in Australasia, Africa and South America.

The survey asked about salaries, job satisfaction, work–life balance, encounters with discrimination, mental health and other key issues that can define and shape a scientific career. The results, along with follow-up interviews with selected respondents, captured the diversity of the scientific experience, from the struggles to the triumphs.

More than two-thirds (68%) of respondents said that they were satisfied or very satisfied with their careers, a rate that is largely unchanged from the 2016 survey. Still, there's no guarantee that those numbers will stay stable. Thirty-seven per cent of respondents said that their satisfaction had worsened in the past year, and just 32% said that it had improved.

A break-down of responses by employment sector showed how attitudes vary across the wide spectrum of scientific paths. Respondents working for non-profit organizations were especially likely to feel satisfied with their jobs (73%), followed closely by respondents in industry (71%), government (68%) and academia (67%). "This supports the fact that there are very fulfilling, high-paying jobs outside of academia," says Susan Porter, dean and vice-provost of graduate and postdoctoral studies at the University of British Columbia in Vancouver, Canada.

Satisfaction numbers have shifted since the 2016 survey, which found a slightly higher proportion of satisfied scientists in academia (65%) than in industry (63%). The differences between the two surveys suggest that the balance between academia and industry has slightly tilted towards companies.

High levels of job satisfaction among researchers have been documented by other surveys, including one earlier this year of PhD holders by the University of British Columbia and one in 2016 of European researchers conducted by Vitae, a non-profit science-career advocacy organization in Cambridge, UK. But Vitae head Janet Metcalfe warns that job satisfaction isn't always a sign of a positive working environment. "Researchers love doing research and therefore can have high job satisfaction, but they can still be experiencing high levels of stress and poor well-being," she says.

SALARY

Questions about salary revealed a deeper divide between sectors. Fifty-nine per cent of respondents in industry said that they were happy with their salaries; by comparison, only 40% of respondents in academia, 41% in non-profits and 49% in government said that they were happy with their pay. Overall, 43% of respondents said that they were happy with their salary. Just over half of all respondents reported a recent pay rise, but it clearly wasn't enough to erase all disappointment.

Sam Proskin, a senior geotechnical engineer with Thurber Engineering in Calgary, Canada is pleased with his salary and career choices. He had hoped to land a job in academia as he finished his PhD in geotechnical engineering at the University of Alberta in Edmonton, Canada, in the mid-1990s. But after testing the academic waters in both the United States and Canada, and finding few opportunities and lots of stories of pressure and competition, he decided that a career in consulting would be a better fit for his skill set and ambitions.

Now, Proskin is in a position to offer advice to other young engineers and geologists who are pondering their futures. He encourages them to keep their options open and avoid the academia-or-nothing mindset. "Maybe it's time to switch that thinking around," he says. "The default mode should be that you are going into industry unless you're very academically inclined."

Mariana Pacheo Blanco, a postdoctoral researcher at BIOCEV, a biotechnology and biomedical academic research centre in Vestec, Czech Republic, is happy with her job in academia. But she also has a major complaint: she's disappointed with the funding opportunities. "I spent five years in Germany where there's a lot of funding for science," says Blanco, who is originally from Mexico but earned her PhD at the University of Munster in Germany. "The difference between Germany and the Czech Republic is considerable."

"The default mode should be that you are going into industry unless you're very academically inclined."

As a postdoc struggling to get by, she's part of another great divide that might be even more fundamental than industry versus academia: the gap between the haves and the have-nots (see 'Different paths'). Just 5% of respondents reported a salary of more than US$150,000 per year. Nearly 30% reported a salary between $50,000 and $80,000, and close to 25% said that they earned between $30,000 and $50,000. At the other end of the scale, 11% reported earning between $15,000 and $30,000, and 12% didn't earn even that much.

Job titles matter when it comes to salary. Although a few professors, managers and research directors reported earning less than $15,000, that end of the salary range was dominated by teachers. About 50% of respondents who said that they are mainly teachers earned less than $30,000, and nearly 30% of research or staff scientists were at the same modest place on the salary scale. The upper ends of the scale were populated mostly by full professors, managers and research directors.

As with the 2016 salary survey, geography proved to be a strong determinant of salaries. Nearly 40% of respondents in Asia reported earning less than $15,000 a year, compared with 2% of respondents in North America. At the top level, 11% of all respondents in North America and Australasia reported earning more than $150,000, putting them far ahead of other regions. Pockets of hardship continue to persist in Europe. Just over 20% of respondents in Europe reported earning less than $30,000 a year compared with just 5% of those in North America — a gap that is unchanged from our 2016 survey.

The survey also reflected gender disparities in salary, especially in scientists who had been in their profession for many years. Among respondents who said that they were in the later part of their career, 33% of men reported earning more than $110,000 a year, but only 23% of women reached that level. For entry level, early-career and mid-career respondents, income brackets were fairly evenly split between the genders. Women were also more likely than men to report being unhappy with their salaries (59% compared with 53%).

The question "Are you happy with your salary?" turned out to be an exercise in relativism. More than 20% of respondents who earned more than $150,000 a year said that they were unhappy with their salaries, whereas 27% of people who earned between $15,000 and $30,000 a year reported that they were happy with their lot, probably because that range fit their expectations and cost of living.

Shrisha Rao, a computer scientist at the International Institute of Information Technology in Bangalore, India, hasn't reached a lofty income bracket, but is happy with his salary and gratified with his work. "I'm fortunate to be at an institution that values me," he says. "I'm not making more than people in the United States, but my income is high for my country and my profession."

JOB PATH

Still, unsurprisingly, academia remains a popular destination: nearly three-quarters (70%) of respondents said that it had been their main goal as they finished their PhDs, which is in line with the aspirations of students who responded to *Nature*'s 2017 graduate-student survey (see *Nature* **550**, 549–552; 2017). Blanco, who studies non-Hodgkin lymphoma, says that she is highly satisfied with her job, mainly because she is finds her topic compelling. "I'm working on cancer, which makes the science more exciting," she says. "It's the hot spot everywhere." But she gives

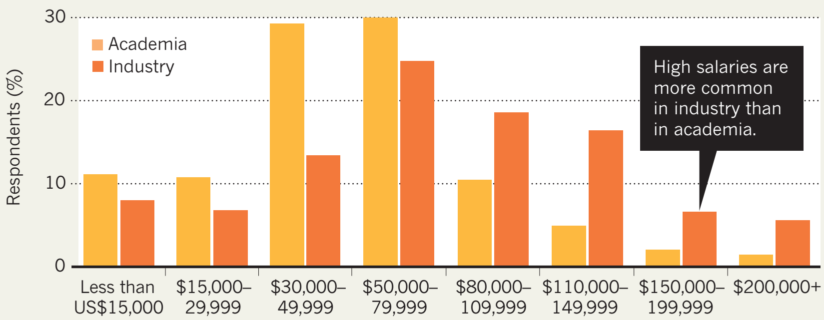
TEREZA CHRBOLKOVA

Mariana Pacheco Blanco enjoys academic research, despite funding struggles.

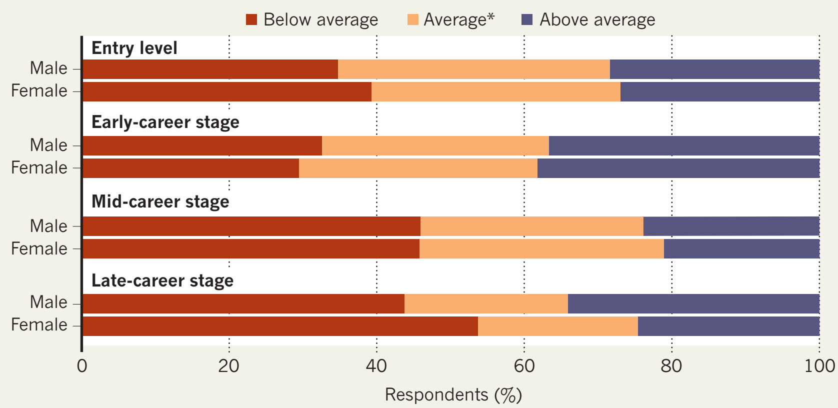
Different paths

Nature's 2018 survey of salary and job satisfaction, which drew more than 4,300 responses from science professionals around the world, found disparities across the science spectrum. Gender, geography and area of employment — academia, industry, government or non-profit — all help to shape career outcomes. Science offers many paths, but some are more rewarding than others.

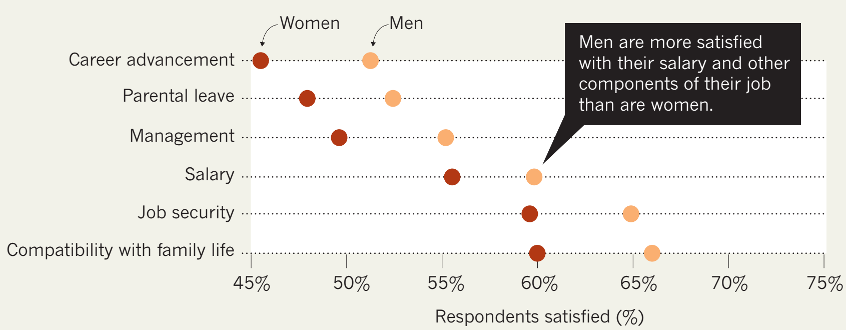
Q What is your current salary?



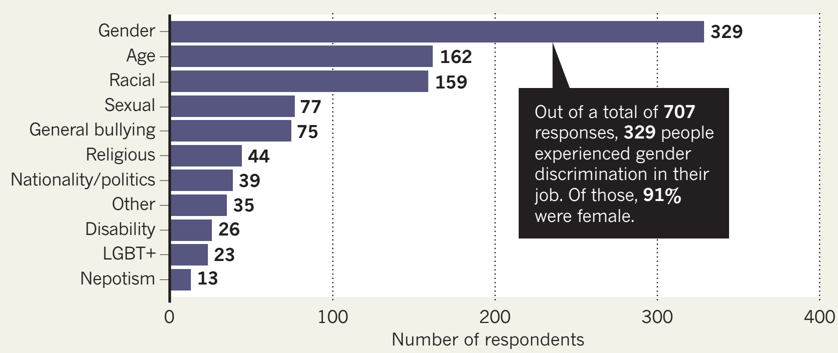
Q Salaries by gender and career stage (all current workplaces)



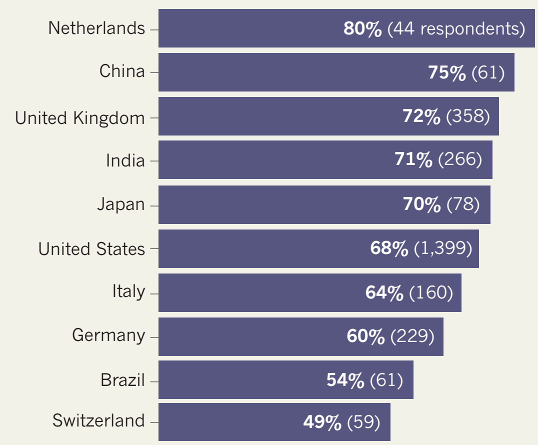
Q Are you satisfied with the following aspects of your current job?



Q Have you experienced discrimination or harassment at your current job?

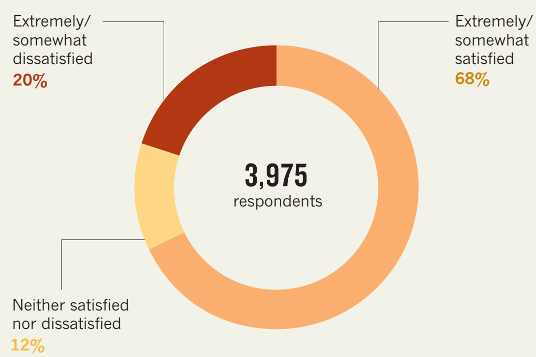


Q Job satisfaction levels in different countries

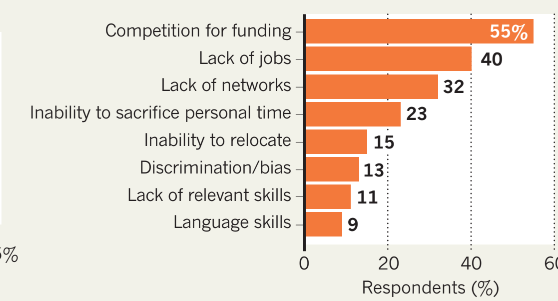


Differences in funding, job opportunities and local politics can make a scientific career more satisfying in some countries than others.

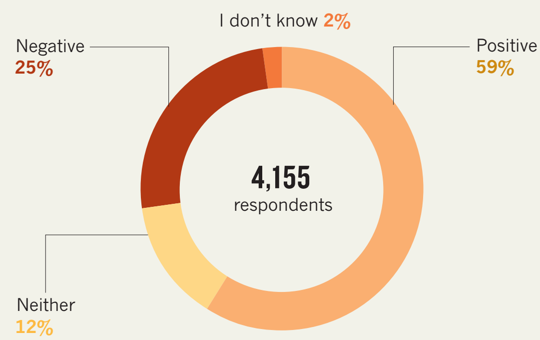
Q How satisfied are you with your current job?



Q What is the biggest challenge to your career progression?



Q Do you have a positive or negative view of your future job prospects?



Some data have been rounded to the nearest per cent and may not add up to 100%. Data-analysis services were provided by educational-research agency Shift Learning. *Average for career stage.

even more credit to her supervisor, cancer biologist Ondrej Havranek, who, is generous with his time and advice, and doesn't demand long hours from his team. "If I want to leave at 5 p.m., I can leave at 5 p.m.," she says. "He respects my private life, which can be a problem for some people in academia."

Rao, who earned his PhD at the University of Iowa in 2005, says that he has had his sights set on academia since the early days of his science training. As a professor at a technology institute where he studies cloud computing and resource utilization, he feels that the plan paid off. "I enjoy a lot of academic freedom," he says. "People aren't telling me what to do on a regular basis." Still, he says, "My peers and I would be happier if we had more resources. India punches below its weight when it comes to science. That's a hard fact."

JENNIFER D. PAN

Aaron Pan switched from the academic track to a curatorial role.

SATISFACTION

Finding enough hours in the day to maintain both a career and a personal life can be a challenge no matter where a scientist works. Although most respondents were satisfied with their own work–life balance, this is another area where industry seemed to have an edge over academic life: 79% of respondents in industry said that they were somewhat or extremely satisfied with this aspect of their career, compared with 68% of respondents in academia.

Job satisfaction is a multifaceted matter. When asked to identify the factor that was most important to satisfaction, respondents put 'interest in the work' at the top of the list. That factor also ranked highest in terms of actual satisfaction — a happy case in which something that's deemed important actually delivers. Still, other aspects of being a scientist are dragging respondents down. Survey participants were generally unsatisfied with their ability to influence decisions that affect them, job security, career-advancement opportunities and recognition for achievements — all things that they felt were important in making a job worthwhile.

WELL-BEING

Respondents were also frank about the negative effects of their work on their mental health. Sixteen per cent of those surveyed said that they had either received help or were currently getting help for depression or anxiety. Seventeen per cent said that they had not received help but would like to. And 3% had sought help but had yet to receive it. Those responses follow a pattern of widespread unease in science. In *Nature*'s 2017 graduate-student survey, 12% of respondents said that they had sought help for anxiety or depression that was directly caused by their studies.

Instances of harassment and discrimination — problems that remain stubbornly common in science — can, of course, also undermine researchers' job satisfaction or ability to perform their job. More than one-quarter (28%) of respondents reported observing such problems at their current job, and more than one-fifth (21%), said that they had personally experienced such treatment. Of those who said that they had witnessed or experienced some sort of discrimination, nearly half (47%), said that they had experienced gender discrimination, the most common type. Ninety-one per cent of respondents who said that they had personally experienced gender discrimination were female. Discrimination based on age (23%) or race (22%) were also relatively common.

About half of all respondents felt that their workplace is doing enough to promote diversity. Those working in industry (58%) were more likely than those in academia (50%) to say that their institution is on top of the issue. Hannah Murfet, a quality compliance manager at Microsoft Research in Cambridge, UK, gives her employer high marks. "Where I work now, the mission is very centred on diversity and inclusion," she says.

Murfet, who is an advocate for women in science, helped to form the Next Generation Network, a group dedicated to helping young researchers learn about careers in compliance and quality control, an increasingly in-demand career option that many scientists fall into almost by accident. "It would be great if more people could consider this opportunity," she says. Most of all, she wants to encourage young people to keep their minds open to a wide range of options. "If you're interested in industry, find an opportunity to get in — it can take you where you want it to," she says. "You don't have to stay as a bench scientist. You can move into marketing, sales or compliance."

> ***"If you're interested in industry, find an opportunity to get in. It can take you where you want it to."***

Nearly 60% of respondents feel positive about their future job prospects — a rate that hasn't changed much from the 2016 survey — but that optimism is not evenly distributed. Scientists were more likely to have a rosy outlook if they had a full-time job and were under the age of 40 and male. The 25% with a negative outlook were more likely to be female and to have a temporary contract. On another pessimistic note, more than half (51%) of respondents said that their job prospects are worse or much worse than they were for previous generations. Still, a full 75% of respondents said that they would recommend a career in scientific research to students — a marked increase from the 61% who took that position in the 2016 survey.

One of those science supporters is Aaron Pan, executive director of the Don Harrington Discover Center, a science museum in Amarillo, Texas. Pan started a postdoc at Southern Methodist University in Dallas, Texas, after earning a PhD in palaeobotany, and was planning to continue on the academic path until another option opened up. He applied for a curator position at the Fort Worth Museum of Science and History in Texas, and his career trajectory changed forever. "I think I would have been happy on many different paths," he says. "Here, I still get to do research, but it's research I want to do. I don't have to publish just to get tenure."

Whether in academia, industry, non-profits or government, there are lots of places to do science, and lots of ways to be a scientist. The *Nature* survey highlights the diversity of options, but it also points to issues that all researchers should keep in mind as they plot their course. From salary to job satisfaction, many things can go right — but they might not. The good news is that science, for most, will always be interesting. And that can be just enough to keep a person going. ■

Chris Woolston *is a freelance writer in Billings, Montana.*

TOTALITY

Keeping it cool.

BY C. L. HOLLAND

Two days ago it had been "we don't negotiate with terrorists", but it was difficult not to negotiate with people who'd stolen the Sun.

In the corner, the TV droned quietly. Most channels weren't broadcasting, but the BBC kept the news running, interspersed with nonsensical cartoons. The breaking-news banner announced that the prime minister, along with other world leaders, had gone to the mothership to negotiate with the Glaosch Imperium, who had turned up to 'invite' Earth to join their empire.

They'd reacted to our (mostly) polite refusal by parking a spaceship the size of the Moon between us and the Sun. Our following, more explosive, refusal had been shrugged off like we were mosquitoes.

"I can't believe you're actually going," Ryan said. "You know what they're doing? Herding us into easily managed groups, is what. You should come with me to the bunker."

I put down my water bottle a little too firmly.

"What else are we supposed to do? The National Grid can't cope with the demand of this constant darkness, and the demand for electricity and gas is only going to increase as it gets colder. You want people dying in their homes because they don't want to put the heating on? We get enough of that when there aren't aliens trying to freeze our asses off. People are hurt and scared. They need a doctor."

Ryan sighed and tugged a hand through his hair. "Fine. I get it. Just promise me if it comes to it, you'll get out and find me."

"I promise."

"And don't bring anyone with you. The bunker's set up for eight, we won't be able to let anyone else in." He pulled me into a hug. "See you, sis."

As I finished packing I wondered how my younger brother's weird hobby of filling our parents' shed with bottled water, candles and tinned food had turned into part-ownership of a nuclear-fallout shelter.

The village hall was packed with camp beds and air beds, most occupied by huddled families under mounds of blankets. There was only so much the ancient heaters could do against the plummeting temperature outside. The darkness already smelt of too many bodies and overworked electrics. I made my way to the pool of light coming from the kitchen to report for duty.

It turned out there wasn't much for me to do. People were quiet, subdued, panic long since burnt out. I passed out plasters and prescribed sleeping tablets and an emergency inhaler, which the village pharmacist dispensed from a suitcase.

Mostly we just watched the news. And the cartoons.

The content was bland and repetitive, obviously designed to stop people panicking again. There was nothing about what would happen to us if the Glaosch didn't move their spaceship. I knew, because Ryan had told me. Before long, most of the plant life would be dead. Animals would die of the cold if starvation didn't take them first. Humans might hang on longer, but since it would be minus 100 degrees in a year, it wouldn't be much longer.

I checked my watch. It said ten past one but I honestly wasn't sure if that was a.m. or p.m.. Or what day it was. The teachers in the crowd were pretty good at keeping routine for the children, but mostly it was just a blur.

There was a wave of shushing and I looked up to see the prime minister on the television, looking even more grey than usual. She was outside Downing Street in the same outfit she'd been wearing when she left.

"After a week of difficult negotiations," she said, "it has been agreed that Planet Earth, henceforth known as Terra, will become a member of the Glaosch Imperium."

There was a collective cry — grief, relief? I couldn't tell. Beside me someone sobbed. "It's over. It's finally over."

You don't know how right you are, I thought. Ryan had seen it coming.

"They don't have to invade us," he'd said. "All they have to do is wait."

The subtitles stuttered out the PM's instructions. "Many of you are gathered safely in school halls and community centres. We ask that you remain there so you can be processed. If you are at home, please stay there; we will find you."

I checked my pocket for my keys and slipped my phone onto the table — Ryan had been very clear about not bringing it with me, or writing down the bunker's location. No one paid any attention as I drifted towards the exit. I grabbed the pharmacy suitcase on the way. It seemed unlikely anyone else would need it now.

Outside, the air felt expectant. I looked up at the sky, feeling a wave of dizziness at the blackness that spread over it like a blanket. No Moon, no stars. *They must be right above us.*

My car was parked out on the road, on double yellow lines, but no one cared about things like that any more. It started first time for a change, as if it knew this was the one that mattered. I threw my satnav out the window. As I drove, the sky got lighter, daylight creeping in like Ryan after curfew, and before long it was like nothing had happened. I fumbled sunglasses from the glove box, even though the Sun was already slipping towards the horizon.

I shivered, and drove into a tomorrow that belonged to the Glaosch Empire. ■

ILLUSTRATION BY JACEY

C. L. Holland *is a British writer of speculative fiction. She has a BA in English with creative writing, an MA in English, and likes to learn things for fun.*

➲ NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

nature

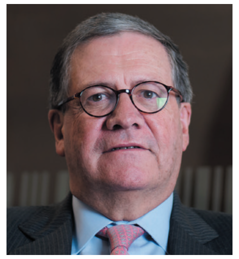
[insideview]

Universidad de los Andes

Profile Feature as seen in *Nature* 25th October 2018

UNIVERSIDAD DE LOS ANDES (UNIANDES): A UNIVERSITY OF SOCIAL INCLUSION AND SUSTAINABLE RESEARCH

A conversation with **PABLO NAVAS**, president of Universidad de los Andes (UNIANDES)



Universidad de los Andes (UNIANDES) is focused on high-quality teaching, social inclusion and sustainable research. Publishing more than 1,000 articles in Scopus each year, its excellence stems from experienced faculty and talented students both at the undergraduate and graduate level. Based in Bogotá, the university is open to collaboration, shown by the proportion of its publications with international networks. University president, Pablo Navas, explains how UNIANDES plans to make Colombia a world-class research destination.

How does UNIANDES innovate in education?
Universidad de los Andes is a liberal, non-profit university that focuses on well-rounded education, with an emphasis on humanistic and scientific thinking. We are in the process of curricular development, focusing on intensive research activities with professors and students, flexibility and interdisciplinarity. Reforms will enhance the connections between teaching and research, promoting students' autonomy. UNIANDES is an increasingly international university with more than 339 undergraduate students and 112 international professors. I believe this is the most important effort we have recently made to improve education.

How is UNIANDES engaged with promoting social inclusion?
We are committed to make UNIANDES the most inclusive university in Colombia. Moreover, our goal is to become a blind-admission university, benefiting from the best minds, regardless of economic and social conditions. We have a vice-presidency role for development and alumni, in charge of fundraising. We are devoted to obtaining funds for talented students who cannot afford our tuition fees. We have created a scholarship program called Quiero Estudiar, based on the principle of reciprocity. The students sign a reciprocity agreement that, in the future, they will help other students to realize the same aspirations they are being helped to achieve.

What are the strategies used by UNIANDES to achieve excellence in research?
We measure our results and impact by international standards. We also strongly believe in research networks, and our most cited publications result from international collaborations. Our office of international affairs identifies strategic partners worldwide with which we promote exchanges.

How does UNIANDES contribute to local or regional issues?
UNIANDES has started to work in different regions of the country. We renewed dialogue with regional universities, NGOs, and local mayors and governors. We aim to collaborate on regional strategic planning. We are also engaged in strengthening our community-based research to understand and respond to local needs. We signed an agreement with the network of national natural parks and are establishing field stations throughout the country. The first one we are creating, and will offer to the research community, is located on the Nevados natural park in the Paramo, a unique Andean ecosystem. These field stations will contribute to sustainable local development through interdisciplinary research.

What have been some of UNIANDES's recent collaborations and networks?
Our most important research network is our most recent one. We have signed an agreement with Pontificia Universidad de Católica Chile and with Tecnológico de Monterrey. We believe that these collaborations, in research, technological innovations and education, will be disruptive and innovative. At a local level, we have signed significant agreements with Agrosavia to contribute to Colombia's agricultural development. Agrosavia is a joint public-private, decentralized, not-for-profit institution for scientific and technical discovery. The corporation's purpose is generating scientific knowledge and developing agricultural technologies. We host three Max Planck Tandem groups, in computational biology, computational physics and international law. We also participate in global initiatives in high-energy physics through CERN and Fermilab.

We host an International Mixed Laboratory, funded by the French Institute IRD (Institut de Recherche pour le Développement) and in collaboration with the Pontificia Universidad Católica de Ecuador (PUCE) to develop collaborative research in the study and protection of Andean ecosystems. We are also very proud to host the South American center for the study of the Sustainable Development Goals, one of the Jeffrey Sachs initiative centers.

What are some examples of impact case studies developed by the university?
We are strengthening our technology transfer strategy. We are developing patents and products that are being transferred to society. Also, some of our innovations have been adopted by several Colombian institutions. For example, the classification of products developed by professors from the schools of Architecture and Design and Arts and Humanities was transferred to Colciencias. The first digital computer in Colombia was installed at UNIANDES, revolutionizing research.

Universidad de los Andes

Springer Nature

nature

[insideview]

Pontificia Universidad
JAVERIANA
Colombia

Profile Feature as seen in *Nature* 25th October 2018

UNIVERSIDAD JAVERIANA: INNOVATIVE RESEARCH FOR PROGRESS AND PEACE

A conversation with **JORGE HUMBERTO PELÁEZ PIEDRAHITA S.J.,** Rector of Pontificia Universidad Javeriana

Universidad Javeriana is one of the largest and most prestigious universities in Colombia. Founded in 1623, it is also one of the oldest educational institutions in South America. Rector Jorge Humberto Peláez S.J. discusses how the research agenda at Universidad Javeriana is contributing to a new Colombian society, with innovative and interdisciplinary approaches that foster collaboration with other institutions.

What makes Colombia an interesting destination for researchers from around the world?
The economic stability in Colombia has favoured the sustained development of high-quality research universities. Additionally, Colombia is exceptionally well-located geographically and is the second most biodiverse country in the world. A recent peace agreement, after 50 years of conflict, has made Colombia a unique peace-building laboratory and has removed barriers for conducting research in a broad range of subjects. We can see from the rising number of researchers coming to Colombia that the country is an interesting hotspot for research during this historic time.

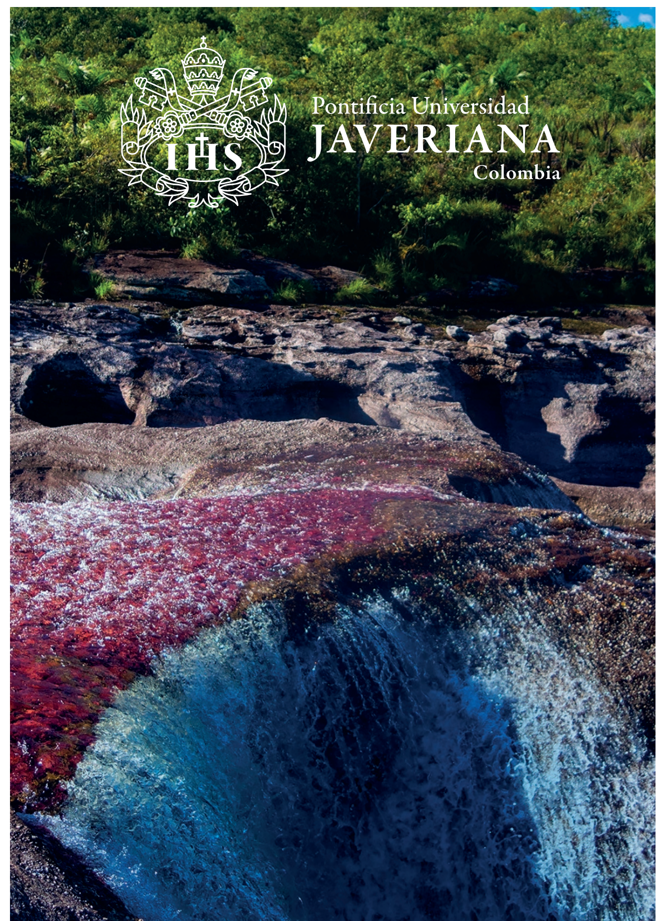
How does Universidad Javeriana contribute to Colombia's development?
The search for innovative answers to local challenges is one of our main contributions to Colombian society. Both research and teaching at Universidad Javeriana play a transformative role. For example, in the field of Health, our Faculty of Medicine studies the mental health impact of the internal conflict on our population. Likewise, in the Social Sciences, the Institute for Intercultural Studies, in our campus in the city of Cali, brings together social, military, business, and former combatant leaders of the FARC — in three strategic conflict zones in southwestern Colombia — to design effective mechanisms to reintroduce participants of the conflict into civilian life. We lead the way in implementing the transformation of the civil war into a constructive dialogue with all members of society. These examples demonstrate that the peace-building process must encompass political, economic, social, and environmental aspects. To this end, Universidad Javeriana has developed strategic alliances with local, national, and international organizations in order to provide local solutions with a global scope.

Furthermore, we highly value inter-institutional collaboration and multidisciplinary approaches to tackle the complex socioeconomic challenges faced by Colombia. For this purpose, our research capacity includes 118 research groups and 12 research institutes encompassing areas such as political science; economics and business; intercultural studies; ecology and territory; public health; clinical and biostatistical epidemiology; infectious diseases; human genetics; immunobiology and cell biology; water and environmental science and engineering; and control systems, power electronics and management of technological innovation.

Since Universidad Javeriana is a Jesuit university, how are ethical values important for promoting research and progress?
Colombia's social and economic progress has been hindered by institutional weaknesses that are reflected, for example, in the country's corruption and deep economic inequality. In many cases, these are a consequence of the absence of sound ethical values. The profound institutional transformation that the country needs to go through calls for creative approaches that will not occur spontaneously. This new environment requires people to be prepared not only with the highest levels of education but also with solid ethics that prioritize human dignity, respect for others, and respect for the common good. As a Catholic and Jesuit university we are committed to educating the whole person, developing individuals who stand out for their values, ethics, academics, and professional quality, as well as a strong sense of social responsibility. It is our mission to promote world-class higher education in order to build a just, sustainable, inclusive, and democratic society.

Could you mention some research highlights that have taken place at Universidad Javeriana in the last few years?
Universidad Javeriana plays a significant role in leading two major research programs of Colombia Científica, the most important initiative funded by the national government to strengthen the quality of Colombian higher education institutions through research. This public funding from Colombia Científica is a recognition of our research achievements. The first program promotes the use of Colombia's biodiversity to produce phytomedicines — medicinal plant-based compounds — to treat cancer. The second program strives to reduce the time for producing and optimizing new sustainable crop varieties by using Omic Sciences (e.g.: genomics, metabolomics, etc.). The development of these programs has fostered international collaboration with leading global institutions, other Colombian universities, and institutions in the private sector. Some of our Colombia Científica program partners include: University of São Paulo, Sorbonne University, Imperial College London, California Institute of Technology, University of Illinois, and University of Tokyo. We are keen on replicating these examples of success in further research collaborations with scholars and institutions from around the world.

Pontificia Universidad JAVERIANA Colombia



32,485 Students
24,203 in Bogotá
8,282 in Cali

118 Research Groups
68 Departments
12 Research Institutes

257 Academic Programs
59 Undergraduate Programs
198 Graduate Programs

We have academic agreements with higher education institutions in
48 Countries

PIGS, SEEDS AND NEW BETS FOR GOOD RESEARCH

In the quiet south of Colombia, far from the capital of Bogotá, researchers from Universidad del Valle are TAKING ADVANTAGE OF LOCAL CREATIVITY AND BIODIVERSITY to establish the Valle del Cauca as a regional centre of development in agricultural and health business.

At Universidad del Valle in Colombia, a group of 20 scientists led by Professor Jose Oscar Gutierrez Montes have successfully completed a pig-to-pig lung transplant using a stem-cell based approach. The recipient, called Pachorrón, received the new lung a year ago and has been living happily since without the need for immunosuppressive medication.

According to Gutierrez "this represents a turning point in transplantation medicine". Indeed, this achievement relies on a recently patented chimaera technology that allows organs to be prepared so that the immune response is suppressed once they are transplanted into the recipient's body.

The researchers have since developed a related approach that could extend these techniques to humans and help to reduce the enormous costs of immunosuppressive therapies that today accompany transplantation.

This accomplishment is the result of more than two decades of multidisciplinary research involving local hospitals and research groups from health and materials engineering schools, fostered by the vice-presidency of research at Universidad del Valle. It follows other achievements, including a recently patented innovation for the preparation of nanostructured photo-sensitive compounds to help fight cancer and an external fixation device for bone fracture reduction that just hit the market in Colombia, Panama and Ecuador.

Today, the vice-presidency for research is building on these achievements by encouraging young scientists and entrepreneurs and follow them with dedicated programs, convinced that high quality research will have a positive impact on the region.

RICH BIODIVERSITY

In equatorial zones, the climate is mild. The countryside is green. Sun and rain alternate. All this makes for a biodiversity hotspot -- and a honeypot for researchers seeking new microbes. One such investigator is David Johnston-Monje, the principal investigator of a Tandem Group with the Max Planck Institute for Plant Breeding Research. This young scientist has pioneered research on endophytes - microbes that are found inside seeds - which were recently shown to have a significant impact on plant growth, fruit size and flavour.

Using biotechnology and metabolomics, the group is tracking plants to study their interactions with microbes, hoping to unveil beneficial organisms. Such research has the potential to improve the competitiveness of the agricultural sector and opens new green business opportunities with super seeds as a means to address climate change and world population growth. For the vice-presidency of research, this new hire will help consolidate the local ecosystem for agricultural research which also includes the recently created Center for Research and Innovation in Bioinformatics and Photonics (CiBioFi) where data analytics and photonics are being used as a predictive tool for sustainable agriculture development.

Universidad del Valle

LOKMAN ILHAN/ANADOLU AGENCY/GETTY

Students and academics march in Bogotá, Colombia, to demand more financial resources for higher education.

Science in Colombia on the cusp of change

The South American country faces deep-rooted problems, but scientists are finding reasons to be hopeful.

BY ALESZU BAJAK

Susana Fiorentino wants to unlock the secrets of Colombian folk medicine. The immunologist, who directs the Immunology and Cellular Biology group at the Pontifical Xaverian University in Bogotá, Colombia's capital of 7 million people, is working with extracts from South America's anamú (*Petiveria alliacea*) and divi-divi plants, which she thinks can be used to treat breast cancer and leukaemia, thanks to their anti-tumour properties.

But while trying to get a phase I trial started with an extract from the divi-divi tree, *Caesalpinia spinosa*, she hit a snag. According to Colombia's drug-safety agency INVIMA, a molecule that identified the extract had to be chemically characterized before the plant could be used in a trial. There were no Colombian laboratories capable of doing the job, so she found labs in China and the United States that could do the work if she posted samples to them. But the request to re-import the fully characterized molecule, to be used in a Colombia-based trial, was denied with no clear reason, says Fiorentino.

The only way to get that potential medicine approved would have been to extract, characterize, synthesize and manufacture it in Colombia, in a laboratory that had not yet been built. Fiorentino had to go back to the plant and isolate another molecule to use as an identifier. She eventually found a laboratory at Colombia's University of Antioquia that could characterize it, instead of having to look abroad for help.

"It was absolutely ridiculous," explains Fiorentino. "Getting that clinical protocol was a path over waterfalls filled with rocks and thorns."

Fiorentino's experience with Colombia's scientific bureaucracy is all too common for researchers in a country emerging from a half-century of civil war. With a peace accord between Marxist rebels and the government signed in November 2016, the country is turning its attention to building the pillars needed to support a strong economy.

Colombia, the only country in South ▸

Q&A
Species seeker

During his PhD at the University of Minnesota in Minneapolis, Juan Fernando Díaz Nieto was part of an expedition that trekked across the tropics of South America and discovered a new species of mouse oppossum (R. S. Voss et al. Am. Mus Novit. **3778***, 1–27; 2013). He's since returned to his native Colombia, where he is a professor of biology at EAFIT University in Medellín.*

What do you study?
I work with marsupials, rodents and bats. I've always been interested in fieldwork — being able to go to unexplored regions that have a high potential for diversity. I also work in museums and use genetic techniques to investigate diversity.

What's it like to do fieldwork in Colombia?
For decades Colombia had zones that were too dangerous to go into. The 2016 peace accord with the FARC, which ended the long civil war here, helped give access to large regions of the country. It's still complicated: to access these areas, one must be careful, work out the logistics and sometimes ask permission — and not just of the government. Many areas have been taken over by other armed groups.

Tell us about a project you recently led.
In July, I and colleagues led a biodiversity project in the Anorí region in Colombia, which is United Nations-backed and involved ex-combatants from the FARC. They weren't serving as guides but as co-investigators. We planned the project with them. After the fieldwork they came to our laboratories in Medellín, working alongside us.

What's the future for science in Colombia?
There's a lot of altruism among researchers here. One person might go into a difficult zone and end up giving the world access to a valuable sample. There are lots of social issues in our country that we have to face. But we are willing — our passion will guide us. ■

INTERVIEW BY ALESZU BAJAK

This interview has been edited for length and clarity.

America to be bordered by both the Caribbean Sea and the Atlantic Ocean, is a mosaic of ecosystems, including mountain ranges, deep jungle, rugged coasts and expansive savannah. It is the second most biodiverse nation in the world (behind neighbouring Brazil) and, thanks in part to this natural laboratory, the country of 49 million people has dozens of universities and institutes working on home-grown science, ranging from the International Center for Tropical Agriculture in Cauca Valley to the natural and physical science laboratories of Bogotá's largest universities (the city has more than 100 tertiary education institutes).

But the key question being asked by Colombian scientists is whether science will find support in the country's post-war economy. For years, their outlook has been pessimistic in the face of strained budgets, meagre resources and red tape that stymie the scientific process, explains Andrew Crawford, a US biologist who, over the course of nine years at Bogotá's Los Andes University, says he has developed a sense of the possibilities and challenges of doing science in Colombia.

A reagent order that could be delivered in a day in the United States or Europe, for example, can take three months in Colombia, says Fiorentino. Monoclonal antibodies, she says, have sat for weeks without refrigeration in customs, useless by the time they reached the lab.

"If we do not improve these administrative problems," Fiorentino stresses, "Colombia will never, never be competitive."

MEAGRE SUPPORT
Many of these problems, researchers point out, are rooted in the government's low financial support for science. Colombia currently invests only 0.67% of its gross domestic product in science and technology, compared with 2.8% in the United States.

Despite this, the country produces on average more than 200 high-quality scientific studies a year on everything from physics to Earth and environmental sciences to the life sciences, according to the Nature Index, which tracks publication in high-quality scientific journals. Recent Colombian papers have made advances in research on dark matter, urban lizards and forest fragmentation. But that output trails behind that of other countries in the region, such as Brazil and Argentina, which until recently invested a lot more in research. Colombia's spending was only one-fifth of Argentina's in 2015, for example. Today, however, uncertain political and economic conditions in those two countries might offer Colombia a chance to catch up.

Last year, 13 Nobel laureates from around the world wrote to then-president Juan Manuel Santos, urging him to raise the government's investment. "The Colombian budget for science and technology for 2018 continues to be extremely low," they wrote, adding: "The consequences will be devastating and irreversible, because science and education are long-term efforts that must be supported consistently."

The letter summarized the growing disenchantment among Colombian researchers about the future of science there. As Enrique Forero, president of the Colombian Academy of Exact, Physical and Natural Sciences in Bogotá, points out, federal funding is decreasing further. "Support from the government is not very acceptable. It's very, very, very low and it seems to be going down every year," he says.

JUAN M. MARTÍNEZ-CERÓN

BRAIN DRAIN
Catalina Pimiento is one Colombian scientist who left and never returned. After graduating from the Pontifical Xavierian University with a degree in biology, she moved first to Mexico, then Panama, the United States, Switzerland, Germany and now Wales, where she has a postdoctoral fellowship at the University of Swansea, investigating mass extinctions like the one that took out the giant shark *Carcharocles megalodon*. As much as it pains her, she won't return to her country of birth.

"Because of my experiences and the career opportunities I've had abroad, I decided to have nothing to do with Colombia," she says. "Colombia does not invest in science. It's that simple."

That's not to say that researchers in Colombia can't make a living — or new discoveries — in the current climate. Juan Fernando Díaz Nieto, a biologist at EAFIT University in Medellín, has a successful career identifying new opossums, rodents and bats in the country's hinterlands that had for decades been inaccessible because of the violence and threat of kidnapping associated with the civil war (see 'Species seeker').

"The peace accord helped with access to large regions of the country," says Díaz Nieto, who recently discovered two new marsupial species along the river of Colombia's Magdalena basin, in the northwest of the country. "We described two new species of mouse opposum that are endemic to Colombia, which we found on either side of the Magdalena River," he explains.

Díaz Nieto, who gained his PhD at the University of Minnesota in Minneapolis, says he was able to make the discoveries because "in lieu of large budgets, we relied on the cooperation of colleagues at other Colombian institutions and cobbled together many small grants from the National Science Foundation, the University of Minnesota and the American Society of Mammalogists."

SCIENCE TAX
The Colombian government is trying to get better at funding science, however. A law passed in 2012 diverted 10% of royalties from natural-resource extraction into a science and technology fund that was allocated across Colombia's 32 government departments. The initiative was designed to help stimulate science by sponsoring research projects and

investing in university research centres in a geographically equitable manner.

Some of it worked. Infectious-disease research was funded along Colombia's Caribbean coast. In Bogotá, Fiorentino received more than US$1 million towards her biopharmaceutical research. And 300 kilometres to the west, a new scientific research centre was built at the University of Caldas in Manizales.

But much of the scheme backfired, in part because of the bureaucracy the new system required. "It was literally impossible to spend the money because of the logistics and paperwork," explains Crawford. "The whole thing was designed upside down."

The nail in the coffin, explains Moises Wasserman, former director of the National University of Colombia in Bogotá, was that approval of these scientific projects was placed in the hands of local politicians instead of scientific review boards. "It was very poorly planned," he says. "The decisions and a large part of the execution were left to very politicized entities who have relatively short tenures — governors with four-year terms."

Under the auspices of provincial governors, millions went to agriculture, aquaculture, energy and infrastructure projects that could hardly be characterized as science, according to an analysis by investigative reporters at La Silla Vacía, a Colombian political news website. The convoluted application process ultimately slowed the disbursement down so much that two of every five pesos in the fund went unspent, the investigation claimed.

Nature approached the Colombian government for comment, but did not hear back before this story went to press.

LOOKING AHEAD

But many Colombian scientists say there's light at the end of the tunnel. Last June, the government proposed a bill that will create a Ministry of Science, Technology and Innovation. The bill is designed to ensure that science has a voice at the highest level of government, according to Iván Darío Agudelo, a senator and the bill's sponsor.

And last July, a month before now-president Iván Duque Márquez was sworn in, he met Forero and 30 other Colombian scientists at the National Academy of Sciences in Bogotá to discuss the future of science in Colombia. By the end of the meeting Duque had pledged, on Twitter, to restructure how science was administered in the country.

"Things seem to be changing," says Forero. "We hadn't had a president in the academy headquarters for 200 years. Or a president-elect. We're not completely invisible any more. People know that we exist."

Some Colombian researchers aren't waiting for the government to increase its investment in science or build a new ministry. Carlos Guarnizo, an ecologist at the Los Andes University, says there are things that individual scientists can do to grow science in the country. Guarnizo and his colleagues started *Ciencia Sumercé*, a science-communication initiative that includes live events at a restaurant in central Bogotá and regularly draws crowds of hundreds, eager to learn about artificial intelligence, climate change or urban ecology.

One priority is keeping the panellists as diverse as he can find. "We want to show that researchers are not just bald old men but ordinary people, of any age, of any gender," says Guarnizo, who hopes his events and accompanying social media activity will draw young people to science.

In that way, he hopes to grow a new generation of science-savvy Colombians. "That could have untold consequences for our country." ■

Aleszu Bajak *teaches journalism at Northeastern University in Boston, Massachusetts. He grew up in Bogotá.*

BIOS COLOMBIA

A science café in Manizales, Colombia, discusses how biotechnology might help local coffee growers.